DNA methylation inheritance in *Arabidopsis*: **The Next Generation**

The work described in this thesis was carried out at the University of Groningen, The Netherlands.

Printed by: RCG Grafimedia - Groningen

ISBN (Print): 978-90-367-9414-5
ISBN (Digital): 978-90-367-9413-8

rijksuniversiteit
groningen

DNA methylation inheritance in *Arabidopsis:* The next generation

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

maandag 28 november 2016 om 9.00 uur

door

Renee Wardenaar

geboren op 5 augustus 1984
te Leek

Promotor

Prof. dr. R.C. Jansen

Copromotor

Dr. F. Johannes

Beoordelingscommissie

Prof. dr. L.H. Franke

Prof. dr. L.W. Beukeboom

Prof. dr. O. Bossdorf

Contents

Abstract

Cytosine methylation is a chemical modification which involves the addition of a methyl group to a cytosine base. Cytosine methylation plays an important role in the regulation of genes and the silencing of transposable elements (TEs) and other repeat sequences. Previous studies have shown that in plants cytosine methylation patterns can be transmitted across cell divisions and even across generations in some instances and contribute to phenotypic variation. DNA methylation is therefore an important epigenetic mark because it contributes to heritable variation in phenotypes that may be of agricultural or evolutionary interest.

With the advent of next generation technologies (e.g. MeDIP-chip; methylated DNA immunoprecipitation followed by hybridization to a tiling array, WGBS-seq; whole-genome bisulfite sequencing) it has become possible to map DNA methylation genome-wide up to a resolution of one base pair. With these new technologies it is now possible to assess the extent to which genome-wide DNA methylation differences can be stably inherited across generations and contribute to phenotypic diversity. However, addressing this question using natural populations is challenging because in natural populations one is confronted with a substantial amount of DNA sequence variation as well. It is difficult to disentangle these two components (sequence variation and variation in DNA methylation) and quantify the contributions of both to phenotypic diversity.

In order to minimize sequence differences, we constructed in *Arabidopsis thaliana* a population of so-called epigenetic recombinant inbred lines (epiRILs; $N > 500$) that were derived from a cross between two parental lines that were near isogenic but have highly divergent methylation patterns (methylomes). One wild-type Col-0 parental line and one mutant Col-0 parental line. The mutant line has lost approximately 70% of its DNA methylation due to a mutation in *DDM1*, which is a gene that is involved in chromatin remodeling.

With the use of the MeDIP-chip technology we reconstructed the methylomes of a subset of the epiRILs ($N = 123$) and the two parental lines. Differentially methylated regions (DMRs) were subsequently detected between the two parental lines and with the use of stably inherited DMRs (Mendelian segregation; $N = 126$) we were able to construct a recombination map for this population. Afterwards we used interval mapping for the detection of epigenetic quantitative trait loci (QTL^{epi}) for two different traits in this population; flowering time and primary root length. Interestingly, the QTL^{epi} explained a high proportion of the broad-sense heritability (~60 - 90 %) of both traits.

Besides studying the contribution of DNA methylation to phenotypic diversity, the epiRILs also provide a good framework to study phenomena that seem to have an epigenetic basis like heterosis. Heterosis describes an F1 phenotype that is superior to its parental varieties (e.g. F1 has more yield). By crossing a subset of the epiRILs (*N* = 19) with wild-type Col-0 we constructed epigenetic F1 hybrids (epiHybrids) which allowed us to study the contribution of epigenetics to heterosis. Several strong heterotic phenotypes were observed among the epiHybrids. Interestingly, heterosis detected for flowering time, leaf area and height could be associated with several heritable parental DMRs. All together our results indicate that epigenetic divergence can also be sufficient to cause heterosis.

Finally, besides induced changes in DNA methylation, stochastic changes in DNA methylation can also be a source of heritable epigenetic and phenotypic diversity in plants. In order to study the dynamics of these stochastic changes we derived robust estimates of the rate at which methylation is spontaneously gained (forward epimutation) or lost (backward epimutation). For this purpose, we used WGBS-seq data from multiple independent *Arabidopsis thaliana* mutation accumulation (MA) lines, which are inbred lines that were propagated for at least 32 generations. The divergence in methylation among these MA lines was compared with those of a large number of *Arabidopsis thaliana* natural accessions, which have diverged from one another for thousands of generations. Our results show that that the dynamic interplay between forward and backward epimutations is modulated by genomic context and show that subtle contextual differences have profoundly shaped patterns of methylation diversity in *Arabidopsis thaliana* natural populations over evolutionary timescales. Besides, theoretical models indicate that the derived epimutation rates are high enough to rapidly uncouple genetic from epigenetic variation, but low enough for new epialleles to sustain long-term selection responses. Altogether these results shed new light on the molecular mechanisms that drive the coevolution of genomes and epigenomes.

Outline of the thesis

In **Chapter 1** an application of the methyl DNA immunoprecipitation (MeDIP)-chip method is described that can be used for the reconstruction of DNA methylomes in *Arabidopsis thaliana*. The chapter presented here describes in detail the MeDIP-chip protocol (wet lab) as well as the subsequent analysis of the hybridization data with the use of a Hidden Markov Model (HMM; bioinformatics). The data analysis steps are separated into four parts involving the preparation of the data, quality assessment and control, implementation of a HMM for the reconstruction of the methylomes and the graphical and biological assessment of the HMM results. Each data analysis step is described and illustrated with the use of an example data set (wild-type Col-0 *Arabidopsis* plant). Several recommendations are made regarding the implementation of the MeDIP-chip protocol and the analysis of the data. The methylomes of the epigenetic recombinant inbred lines (epiRILs) and parental lines were reconstructed with the approach described in this chapter.

In **Chapter 2** the MeDIP-chip data of two epiRILs are evaluated in the context of whole-genome bisulfite sequencing (WGBS-seq) data. The chapter presented here demonstrates and illustrates that when a large number of methylomes need to be reconstructed (e.g. epiRIL study) and high resolution measurements are not needed, the lower resolution and easier implementable (and currently also cheaper) array-based MeDIP-chip technology might be favorable over the single base pair resolution WGBS-seq technology. Three different implementations of the MeDIP-chip technology (differential labeling and dye-swap) were benchmarked against WGBS-seq and compared among each other. Results showed that MeDIP-chip performed reasonable well when appropriate data preparations steps were taken and the appropriate analysis tools were applied. Based on the results, several recommendations are made regarding the implementation of the MeDIP-chip technology for the analysis of DNA methylation in *Arabidopsis*.

In **Chapter 3** the combined effect of removing sequence variation and DNA methylation on the meiotic recombination landscape of an *Arabidopsis* mapping population is tested. With the use of the reconstructed methylomes (Chapter 1) several differentially methylated regions (DMRs) could be detected between the parental lines (wild-type and *ddm1-2* mutant parent). A recombination map was constructed with the use of an informative subset of these DMRs that segregated in a Mendelian fashion in the epiRIL population (stable DMRs). Estimated genetic

lengths of each chromosome indicated that the recombination rates are, on a global scale, comparable with those of classical *Arabidopsis* crosses. However on the local scale we demonstrate that the recombination rate is decreased at the boundaries of pericentromeric regions and increased in chromosomal arms. Despite the fact that recombination barriers have been forced to a minimum, it becomes clear that the differences in recombination are not strong enough to place the epiRILs outside of the natural range.

In **Chapter 4** the recombination map described in Chapter 3 is used in combination with classical linkage analysis to search for epigenetic quantitative trait loci (QTL^{epi}) underlying complex traits in the epiRIL population. Using interval mapping highly significant QTL were detected for flowering time and primary root length. The results could be confirmed with data from replicate phenotyping experiments. The combined additive effects of the QTLs of both traits explained a substantial proportion of the broad-sense heritability. Furthermore analysis of more than hundred natural accessions showed that a high proportion of the experimentally induced DMRs were also variable in nature. This observation indicates that these DMRs could also function as QTL^{epi} in nature and thus constitute a measurable component of the so-called “missing heritability”.

In **Chapter 5** we explore whether epigenetic divergence is sufficient to trigger heterosis in *Arabidopsis thaliana*. To do so we created epigenetic F1 hybrids (epiHybrids) by crossing Col-0 wild-type as a maternal parent to 19 near-isogenic *ddm1-2*-derived epigenetic recombinant inbred lines as paternal parents (epiRILs; Chapter 3). We monitored several traits such as leaf area, flowering time, final plant height, rosette branching, main stem branching and growth rate, and observed that the majority of the epiHybrids exhibited heterotic phenotypes for at least one of these traits. Furthermore by treating mid-parent divergence (the divergence from the mid-parent value; mean phenotypic value of both parents) as a phenotype we were able to detect significant QTL for leaf area, flowering time and plant height. The variation in mid-parent divergence could not be explained by the presence of TE-associated structural variants which indicates the QTL most likely have an epigenetic basis. Altogether these results indicate that epigenetic variation, like variation in DNA methylation, can contribute to heterosis independently from DNA sequence variation. Our findings might therefore have implications for future crop breeding.

In **Chapter 6** robust estimates are derived for the rate at which methylation of individual CG cytosines is stochastically gained (forward epimutation) or lost

(backward epimutation) in the model plant *Arabidopsis thaliana.* CG methylation divergence was calculated as the proportion of differentially methylated CG cytosines between individuals of independent mutation accumulation (MA) lines. A model was developed to quantify the CG methylation divergence as a function of the divergence time between the individuals and the forward and backward epimutation rates. By analyzing different annotations (e.g. genes, transposable elements, etc.) we demonstrate that the dynamic interplay between forward and backward epimutations is context specific and that this specificity has shaped the patterns of methylation divergence in *Arabidopsis*. Furthermore, theoretical models indicate that the epimutations rates are high enough to uncouple genetic variation from epigenetic variation but that the rates are low enough, for new epialleles to sustain long term selection responses.

Chapter 1

Genome-wide analysis of DNA methylation in *Arabidopsis* using MeDIP-chip

---

Published as:

Cortijo S*, Wardenaar R*, Colomé-Tatché M, Johannes F, Colot V (2014) Genome-wide analysis of DNA methylation in *Arabidopsis* using MeDIP-chip. *Methods Mol Biol* **1112**:125-149.

*Equal contribution

Chapter 1

Abstract

DNA methylation is an epigenetic mark that is essential for preserving genome integrity and normal development in plants and mammals. Although this modification may serve a variety of purposes, it is best known for its role in stable transcriptional silencing of transposable elements and epigenetic regulation of some genes. In addition, it is increasingly recognized that alterations in DNA methylation patterns can sometimes be inherited across multiple generations and thus are a source of heritable phenotypic variation that is independent of any DNA sequence changes. With the advent of genomics, it is now possible to analyze DNA methylation genome-wide with high precision, which is a prerequisite for understanding fully the various functions and phenotypic impact of this modification. Indeed, several so-called epigenomic mapping methods have been developed for the analysis of DNA methylation. Among these, immunoprecipitation of methylated DNA followed by hybridization to genome tiling arrays (MeDIP-chip) arguably offers a reasonable compromise between cost, ease of implementation and sensitivity to date. Here we describe the application of this method, from DNA extraction to data analysis, to the study of DNA methylation genome-wide in *Arabidopsis*.

1.1 Introduction

In eukaryotes, DNA methylation almost exclusively affects cytosines (5-methylcytosines). Once established, this modification can be maintained over numerous cell divisions and even across generations in some instances. However, it remains unclear to what extent differences in DNA methylation can be stably inherited and this question is the subject of intense studies. This is especially true in *Arabidopsis*, where epigenetic recombinant inbred lines (epiRILs) have been derived from parents with few differences in DNA sequence but contrasted DNA methylation profiles [1, 2]. One such population of epiRILs has been epigenotyped [3] in order to assess the stability of parental DNA methylation differences and their impact on several complex traits. Here, we describe the methyl DNA immunoprecipitation (MeDIP)-chip protocol used to reconstruct the DNA methylome maps, starting from the extraction of DNA to analysis of the hybridization data using Hidden Markov Models (HMM; *see* flow chart in Fig. 1). Subheading 1.2 lists the materials needed for the “wet” part as well as the software and data used for analysis. Subheading 1.3 describes step by step the MeDIP-chip experiment (Subheadings 1.3.1 and 1.3.2) and the analysis of hybridization data, starting from data preparation (Subheading 1.3.3), then quality assessment and control (Subheading 1.3.4), implementation of a HMM

for reconstructing DNA methylome maps (Subheading 1.3.5) and graphical and biological assessment of HMM results (Subheading 1.3.6).

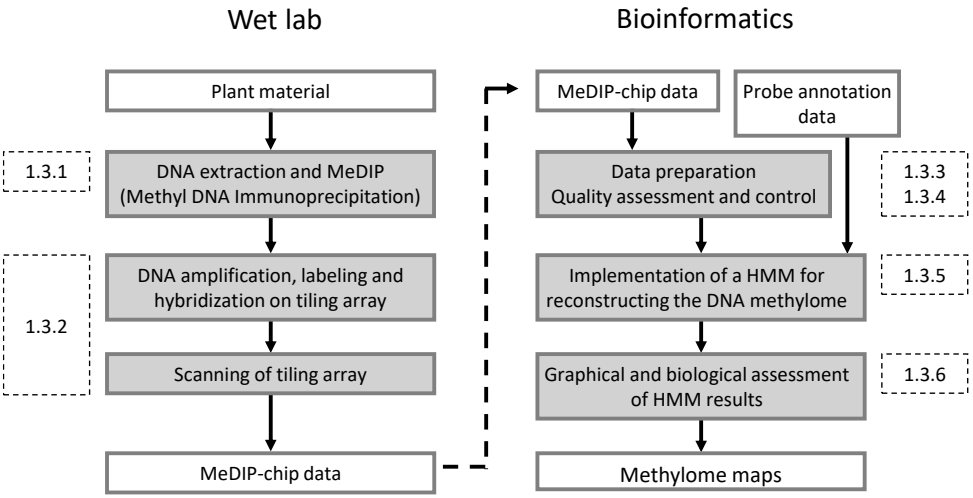

Figure 1. Flowchart for the reconstruction of methylome maps (Subheading 1.3).

1.2 Materials

1.2.1 DNA extraction and methyl DNA immunoprecipitation

1. DNA extraction with DNeasy plant Maxi kit (Qiagen, Catalogue N° 68163).
2. 1.5 mL Siliconized tubes: Clear-view™ Snap-Cap microtubes, siliconized (Sigma, Catalogue N° T4816-250EA).
3. Sonicator: Bioruptor (Diagenode, Catalogue N° UCD-200).
4. Buffer 1: 13.3 mM Tris HCl pH 7.5, 667 mM NaCl, 1.3 mM EDTA.
5. Monoclonal antibody against 5mC (Diagenode, Catalogue N° MAb-006-500).
6. Rotating wheel.
7. Magnetic beads: M280 Dynabeads (Invitrogen, Catalogue N° 112-01D).
8. Buffer 2: 10 mM Tris HCl pH 7.5, 500 mM NaCl, 1 mM EDTA.
9. Buffer 3: 30 mM Tris HCl pH 8.0.
10. Proteinase K: 20 µg/µL (NEB, Catalogue N° P8102S).
11. Phenol/chloroform/IAA (25:24:1, pH 8.0) and Chloroform/IAA: (24:1).

12. Glycogen azure: 20 µg/µL, resuspended in water (Sigma, Catalogue N° G5510-1G).
13. NaOAc: 3 M pH 5.2.
14. MinElute Reaction Cleanup Kit (Qiagen, Catalogue N° 28204).
15. PicoGreen: Quant-it PicoGreen dsDNA reagent (Invitrogen, Catalogue N° P7581) diluted to 0.5 % in TE pH 8.0.

1.2.2 DNA amplification, labeling and hybridization on tiling array

1. WGA2 kit (Sigma, Catalogue N° WGA2-50RXN).
2. QIAquick PCR Purification Kit (Qiagen, Catalogue N° 28104).
3. Dual Color DNA labeling kit (NimbleGen, Catalogue N° 06370250001).
4. Hybridization and wash buffer kits (NimbleGen, Catalogue N° 05583683001 and 05584507001).
5. Scanner: High-Resolution (2 µm) Microarray Scanner (Agilent, Catalogue N° G2565CA).
6. NimbleScan software (NimbleGen).

1.2.3 Software requirements

This protocol requires R for the analysis of the MeDIP-chip data. R is a command line-based software environment for statistical computing and graphics. It can be freely downloaded at http://www.r-project.org and installed on all three main operating systems (Windows, Unix/Linux and Mac). Instructions about installation and tutorials can be obtained from the same website. R is extensively used among biostaticians due to the availability of statistical packages for the analysis of a broad spectrum of biological data. In addition to R, we also recommend downloading a text editor with syntax highlighting (e.g. Notepad++). Programming mistakes are more easily detected when using a text editor. All the code lines and functions are highlighted throughout the chapter in `courier font`. The HMM is implemented in C++. An electronic version of the R code presented in this chapter and the HMM software are freely available at the following URL: http://www.johanneslab.org/publications. This chapter does not show the code for generating the figures. This code can, however, be downloaded from the same URL.

1.2.4 Dataset

The protocol was implemented for the efficient and cost-effective genome-wide study of DNA methylation of a large number of *Arabidopsis* lines. The dataset used to illustrate this protocol can be downloaded from the above URL and consists of six files that contain the measured signal intensities (IP and INPUT) for one wild-type line (Columbia accession, Col-0), probe annotation, conservation scores for probes and an example of an array with a hybridization artefact.

1.2.4.1 Methylation data

The methylation data should be tab-delimited and have the format shown in Table 1. The first column of the file should contain the probe identifier and the remaining column (or columns when replicates are available) should contain the measured probe intensities. The IP and INPUT files should have the same tab-delimited format.

Table 1. Format methylation data.

PROBE_ID	REP1_INPUT_RED	REP2_INPUT_RED	REP3_INPUT_RED
CHR01FS000000061	778.53	2534.67	1033.31
CHR01FS000000212	2366.51	2756.02	1333.69
CHR01FS000000382	4028.27	7776.75	3201.88
CHR01FS000000507	13685.61	15014.29	8556.37
CHR01FS000000707	1565.45	2626.51	1187.04

1.2.4.2 Hybridization artefact data

For illustrative purposes we also show an example of a hybridization artefact (Fig. 2a). This file should also be tab-delimited and have the format shown in Table 2. The first column should again contain the probe identifier, the second and third column should contain the location of the probe on the array (x and y position on the array) and the fourth column (PM) should contain the measured probe intensity (IP or INPUT signal).

Table 2. Format hybridization artefact data.

PROBE_ID	X	Y	PM
CHR01FS000000061	327	1335	3219.96
CHR01FS000000212	191	1257	4840.31
CHR01FS000000382	826	34	8668.02
CHR01FS000000507	731	529	19781.76
CHR01FS000000707	624	562	1195.29

1.2.4.3 Conservation score data

The conservation score of a probe indicates the uniqueness of the probe sequence (not all probe sequences are unique). This score was obtained by performing a BLAST search. Scores are percentage of identity with the second best hit (the first hit is the location in the genome for which the probe was designed). Probes can be visualized at http://epigara.biologie.ens.fr/index.html. The conservation score data should have the tab-delimited format shown in Table 3.

Table 3. Format conservation score data.

PROBE_ID	SCORE
CHR01FS000000061	73
CHR01FS000000212	56
CHR01FS000000382	64
CHR01FS000000507	62
CHR01FS000000707	74

1.2.4.4 Annotation data

The annotation files contain the probe identifiers of probes that are located within introns of protein coding genes or transposons. The annotation data should only contain one column with probe identifiers as shown in Table 4.

Table 4. Format annotation data.

PROBE_ID
CHR01FS000004351
CHR01FS000005311
CHR01FS000007129
CHR01FS000007479
CHR01FS000007814

1.3 Methods

1.3.1 DNA extraction and MeDIP

1. Extract DNA from plant material (1–2 g fresh weight, we use aerial parts of three-week-old plants grown under long day conditions) with Qiagen DNeasy plant Maxi kit. 1.8 µg of DNA is needed for this protocol (includes sonication test and INPUT and IP fractions).
2. Quantify DNA and place 1.8 µg in a final volume of 180 µL (complete with water if necessary) in 1.5 mL siliconized Eppendorf tubes. Set aside 2 µL (corresponding to 20 ng of DNA) for sonication control. Sonicate the remaining 178 µL using seven cycles of 30 s ON/30 s OFF. Note that all six positions within the sonicator need to be filled with an equal volume of water (178 µL) put in each tube. Place all six tubes in an ice bucket and add ice to the sonicator bath to cool it off. Repeat sonication once (14 cycles in total). Keep 13 µL to test sonication (sonicated fraction).
3. Run non-sonicated (2 µL) and sonicated (13 µL) samples side by side in 1.5 % 1 × TAE gel. A smear should be visible between 100 and 600 bp (with maximum intensity around 300 bp) after sonication.
4. Keep 15 µL to serve as INPUT (150 ng). Use the remaining 150 µL of sonicated DNA (1.5 µg) for IP.
5. Add 450 µL of buffer 1 to IP fraction (total volume of 600 µL). Incubate 10 min at 95 °C to denature DNA (this is critical as the antibody only recognizes 5mC on single-stranded DNA!) and let sit on ice for 2 min. Add 5 µL of 1 µg/µL anti-5mC antibody to denatured IP fraction. Close tubes, wrap with parafilm

(siliconized tubes tend to leak) and incubate overnight at 4 °C with gentle agitation (we use a rotating wheel, with a 45° inclination, 8 rpm).

6. Use 40 µL of magnetic beads per MeDIP. Prepare a tube with the total amount of beads required for the number of MeDIP performed. Wash the beads three times with 1 mL of buffer 2 (*see* **Note 1**) and resuspend one last time with buffer 2 in the starting volume. Put 40 µL of washed beads (make sure that they are well resuspended by pipetting up and down the slurry several times) into each MeDIP tube. Put on rotating wheel for 4 h at 4 °C with gentle agitation (45° inclination, 8 rpm).
7. Put IP samples on the Dynabeads rack (magnetic rack). Collect supernatant in a new 2 mL Eppendorf tube (supernatant fraction). Add 300 µL of buffer 2 to IP tube. Agitate briefly by hand and place for 10 min at room temperature on the rotating wheel with gentle agitation (45° inclination, 8 rpm). Put back on the Dynabeads rack and add first wash to supernatant fraction. Perform three more washes, each time with 600 µL of buffer 2. Discard washes.
8. Add 300 µL of buffer 3 to the IP pellet after last wash and transfer IP and supernatant fractions to 1.5 mL and 2 mL tubes, respectively (*see* **Note 2**). Add 7 µL of Proteinase K to elute. Incubate 1 h at 42 °C, with occasional shaking.
9. Add one volume of phenol/chloroform/IAA to the IP and supernatant fractions (300 and 900 µL, respectively). Vortex and centrifuge 5 min at 14,000 × g at room temperature. Place aqueous phase (top phase) in a new tube. Add one volume of chloroform/IAA to aqueous phase. Vortex and centrifuge 5 min at 14,000 × g at room temperature. Place aqueous phase in a new tube.
10. To precipitate DNA, add 1 µL of glycogen azure, 1:10 volume of NaOAc and one volume of isopropanol to the IP and supernatant fractions (30 and 90 µL for NaOAc and 300 and 900 µL for isopropanol in IP and supernatant fraction, respectively). Vortex between additions of each component. Keep at −20°C for at least 1 h or overnight. Centrifuge 30 min at room temperature at max speed (>13,000 × g). Discard supernatant and add 500 µL of ethanol 70 %. Mix and centrifuge for 20 min at room temperature at max speed (>13,000 × g). Discard the supernatant and dry DNA pellets by leaving the tubes open on the bench for ~30 min. Resuspend all DNA pellets in 40 µL of TE pH 8.0 and add 25 µL to INPUT fraction.
11. Perform quantitative PCR on the three fractions (IP, supernatant and INPUT) with known positive and negative controls before proceeding with purification, labeling and hybridization to tiling array. Note that for wild-type

Arabidopsis (Columbia accession) approximately 10–20 % of the genome should be immunoprecipitated with the anti-5mC antibody for DNA extracted from aerial or root parts.

12. DNA should be cleaned one last time using the MinElute kit (*see* **Note 3**). Expect 30 % loss of DNA.
13. DNA concentration is checked with NanoDrop 3300. Add 2 µL of diluted PicoGreen at 0.5 % to 2 µL of DNA and quantify this mix using function "dsDNA PicoGreen® dye" in "Nucleic Acid Quantitation" (*see* **Note 4**).

1.3.2 DNA amplification, labeling and hybridization on tiling array

1. Use 10 ng of IP and 50 ng of INPUT fractions for amplification with the WGA2 kit. Start from the "Library preparation" step of the protocol, as there is no need for the DNA fragmentation step.
2. Purification of the amplification products is carried out using QIAquick PCR Purification Kit. Quantify and run in a 1.5 % agarose 1 × TAE gel. This should produce a smear corresponding to the sonication smear (between 100 and 600 bp). Final yield fluctuates between 3 and 6 µg.
3. DNA labeling is carried out using the Dual Color DNA labeling kit, using 1 µg of amplified IP and INPUT DNA. Resuspend labeled DNA in 20 µL of water and quantify it, together with Cy3 and Cy5 using the "microarray function" of the NanoDrop 2000. One should expect 10–20 µg of DNA after labeling and 200–400 pmol of incorporated dye. Repeat labeling if DNA yield or incorporation levels are less than 5 µg or 100 pmol, respectively (*see* **Note 5**).
4. Differential hybridization is carried out using a NimbleGen 3x720K tiling array design (three identical chambers, design available on request) and following the manufacturer's instructions. Use 4 µg of each of the two labeled DNA samples (IP and INPUT) per chamber. Hybridization is in dye-swap (IP in red and INPUT in green for the first chamber and vice versa for the second chamber).
5. After washing, the NimbleGen 3x720K tiling array is scanned using a High-Resolution (2 µm) Microarray Scanner (Agilent). It is preferable to scan each chamber independently.
6. Grid alignment and pair files extraction are made using the NimbleScan software and following the manufacturer's instructions.

1.3.3 Data preparation

Following the "wet lab" part one is confronted with a substantial amount of data ready to be analyzed. Before we show how this can be achieved, we detail several data preparation steps. The following commands are used to import the data in the R workspace. The command `setwd()` sets the working directory, such that there is no need to define the complete pathname of your files. The command `head()` shows the first lines of the file.

```
> setwd("D:\\reconstruction_methylome_maps")
> input_wt <- read.table(file="input_wild_type.txt",
+ header=TRUE,sep="\t")
> ip_wt <- read.table(file="ip_wild_type.txt",
+ header=TRUE,sep="\t")
>
> head(input_wt)
         PROBE_ID REP1_INPUT_RED REP2_INPUT_RED REP3_INPUT_RED
1 CHR01FS000000061         778.53        2534.67        1033.31
2 CHR01FS000000212        2366.51        2756.02        1333.69
3 CHR01FS000000382        4028.27        7776.75        3201.88
4 CHR01FS000000507       13685.61       15014.29        8556.37
5 CHR01FS000000707        1565.45        2626.51        1187.04
6 CHR01FS000000827        5939.94        7285.02        3212.73
  REP1_INPUT_GREEN REP2_INPUT_GREEN REP3_INPUT_GREEN
1           408.61          2038.57           818.98
2           712.76          2019.65           649.84
3          1350.67          5406.18          2090.43
4          2980.53          9570.41          5614.20
5           611.33          2405.53           460.63
6          1162.24          4555.31          2311.96
```

The IP and INPUT data have the same format; hence, there is no need to show the first lines of both files. We convert the data to a logarithmic scale using the following commands:

```
> log2_ip_wt              <- log2(ip_wt[,2:7])
> log2_ip_wt              <- data.frame(ip_wt[,1],log2_ip_wt)
> names(log2_ip_wt)[1]    <- "PROBE_ID"
>
> log2_input_wt           <- log2(input_wt[,2:7])
> log2_input_wt           <- data.frame(input_wt[,1],log2_input_wt)
> names(log2_input_wt)[1] <- "PROBE_ID"
```

After log transformation the datasets will have the same format only the signal intensities will be log transformed. In order to determine enrichment for DNA

methylation one has to calculate the intensity ratio of the IP and INPUT signal (log_2 ratios). The following commands are used to calculate the intensity ratios. The dye-swapped replicates have been treated separately in this case (i.e., IP_{green} and $INPUT_{red}$ and vice versa). The IP and INPUT signals have also been averaged.

```
> wt_ip_green    <- (log2_ip_wt[,5]+log2_ip_wt[,6]+log2_ip_wt[,7])/3
> wt_input_red   <- (log2_input_wt[,2]+log2_input_wt[,3]+
+ log2_input_wt[,4])/3
> wt_green_red   <- wt_ip_green-wt_input_red
> wt_green_red   <- data.frame(log2_ip_wt[,1],wt_green_red)
> names(wt_green_red) <- c("PROBE_ID","GREEN_RED")
>
> wt_ip_red      <- (log2_ip_wt[,2]+log2_ip_wt[,3]+log2_ip_wt[,4])/3
> wt_input_green <- (log2_input_wt[,5]+log2_input_wt[,6]+
+ log2_input_wt[,7])/3
> wt_red_green   <- wt_ip_red-wt_input_green
> wt_red_green   <- data.frame(log2_ip_wt[,1],wt_red_green)
> names(wt_red_green) <- c("PROBE_ID","RED_GREEN")
```

After the calculation of the intensity ratios the dye-swap signals can be calculated using the following code:

```
> wt_dye_swap <- (wt_green_red[,2]+wt_red_green[,2])/2
> wt_dye_swap <- data.frame(wt_green_red[,1],wt_dye_swap)
> names(wt_dye_swap) <- c("PROBE_ID","DYE_SWAP")
```

The dye-swap should account for possible dye bias in experiments. The data is now ready for subsequent analysis steps.

1.3.4 Quality assessment and control

Prior to array data analysis, we conduct detailed quality checks of each tiling array experiment. This quality assessment is necessary to ensure biologically meaningful results later on. If the data contains systematic hybridization artefacts or technical variation beyond a certain acceptable level it is advisable to remove or to repeat the bad sample. We distinguish between two levels of quality assessment. The first level relates to the quality of the overall hybridization experiment and the second level to the quality of the individual probes.

1.3.4.1 Quality of the overall hybridization experiment

We start by evaluating the distribution (or spreading) of the DNA fragments over the tiling array. This can be achieved by visual inspection of the array image within each separate channel (Fig. 2a). By design, the signals should be randomly distributed and show no systematic spatial patterns. Artefacts such as scratches and bright spots can be easily detected in this way. The following commands are used to import the data in the R workspace:

```
> hybr_artefact <- read.table(file="hybridization_artefact.txt",
+ header=TRUE,sep="\t")
> head(hybr_artefact)
          PROBE_ID   X    Y        PM
1 CHR01FS000000061 327 1335  3219.96
2 CHR01FS000000212 191 1257  4840.31
3 CHR01FS000000382 826   34  8668.02
4 CHR01FS000000507 731  529 19781.76
5 CHR01FS000000707 624  562  1195.29
6 CHR01FS000000827 927  485  7460.27
```

Plotting the reconstructed array image involves log transformation of the measured signals (PM) and rescaling of the log transformed signal between 0 and 1 in order to convert the signal into RGB colors. The code for plotting the array image (Fig. 2a) can be found at the above URL (*see* end of Subheading 1.2.3).

We also evaluate whether a sufficient amount of DNA was hybridized to the array. This can be done by plotting the density of the signal of each separate channel (Fig. 2b). The detection range of the signal has a lower and upper bound. In the case of insufficient DNA, there will be a rapid increase of probe signals in the lower detection range. Conversely, in the case of too much DNA the signal distribution will become saturated in the upper detection range. Both scenarios can seriously compromise the sensitivity of the technology to capture biologically meaningful variation. To illustrate this, we plot the density of the input signal of two different arrays (Fig. 2b) using the plotting code that is provided as a text file (*see* end of Subheading 1.2.3).

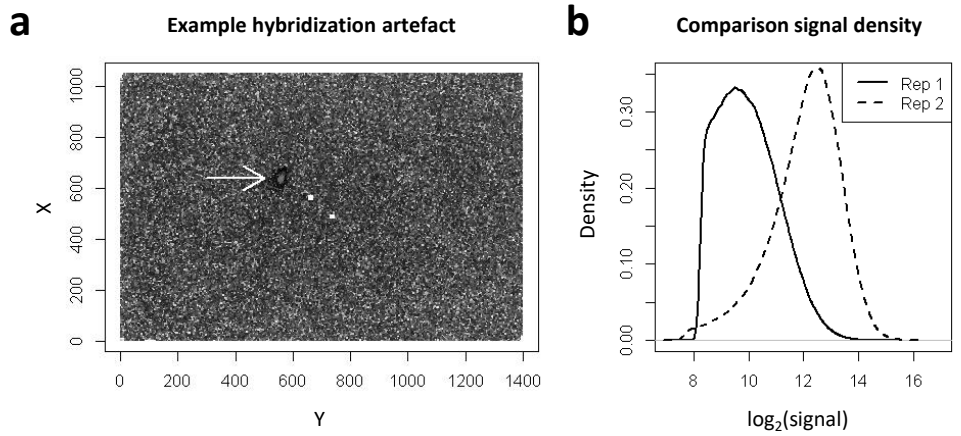

Figure 2. Quality of the overall hybridization experiment. (**a**) The arrow points to an unwanted spatial artefact on the tiling array. One could consider excluding the relevant probes or discarding the tiling array entirely. (**b**) Shown is the signal density distribution of the Cy3 INPUT channel for two replicates. One of the individuals (solid line) shows a steep increase in the lower signal range suggesting that an insufficient amount of DNA was hybridized to the tiling array. The signal distribution of the other replicate (dashed line) is normal. The bulk of the data is located in the center of the detection range indicating that the right amount of DNA was hybridized to the tiling array.

1.3.4.2 Quality of individual probes

The second quality assessment level is the quality of the probes. NimbleGen arrays are designed to minimize cross-hybridization as much as possible. However, given the large number of probes and near full genome coverage, it is difficult to exclude possible cross-hybridization events. Such events occur when non-target sequences hybridize with probes on the array, leading to exaggerated signal intensities. It may therefore be desirable to identify probes, *a priori*, that have multiple similar or exact matches in the genome. We assess this by calculating the so-called conservation score. This score is obtained by performing a BLAST search. Scores are percentage of identity with the second best hit (the best hit is the location on the genome for which the probe was designed). We decided to flag probes that have a conservation score higher than 85 (Fig. 3a). For simplicity we here provide a complete dataset with conservation scores already assigned to each probe. This data can be inputted as follows:

```
> cons_score_probes <- read.table(file="conservation_score.txt",
+ header=TRUE,sep="\t")
> head(cons_score_probes)
          PROBE_ID SCORE
1 CHR01FS000000061    73
2 CHR01FS000000212    56
3 CHR01FS000000382    64
4 CHR01FS000000507    62
5 CHR01FS000000707    74
6 CHR01FS000000827    66
```

One can use the plotting code which is provided as a text file to plot the density histogram of the conservation scores of the probes as shown in Figure 3a.

In addition to the above *a priori* screening of potential cross-hybridizing probes we utilize another quality criterion, which involves assessing the consistency of probe signals for the INPUT across biological or technical replicates (provided they are available). To do this, we identify a probe's signal rank in the overall array signal distribution of one replicate array and compare it to its rank in the distribution of the other arrays. Inconsistent probe signals will show large variation in ranks and should be treated with caution. If we consider the three dye-swapped biological replicates (3 × 2 arrays) of the Col-0 accessions, there are six rank values for each probe and we can calculate its rank variance. Doing this for each probe on the array yields a rank variance distribution, which can be used to spot outlying probes (Fig. 3b). For example, we may want to consider excluding or flagging probes with a rank variance of more than 3 standard deviations from the mean (Fig. 3b). We use the following code to determine the rank and the rank variance of the probes as well as the three standard deviation cutoff:

```
> probe_rank <- apply(log2_input_wt[,2:7],MARGIN=2,rank)
> determine_rank_var <- function(x){
+    mean_val   <- mean(x)
+    mean_dif   <- abs(x-mean_val)
+    extreme    <- which(mean_dif == max(mean_dif))
+    sd_ext     <- sd(x[-extreme])
+    return(sd_ext)
+ }
> rank_var   <- apply(probe_rank,MARGIN=1,determine_rank_var)
> rank_var   <- data.frame(log2_input_wt[,1],rank_var)
> names(rank_var) <- c("PROBE_ID","SD")
> mean_var   <- mean(rank_var[,2])
> sd_var     <- sd(rank_var[,2])
> sd_cutoff  <- mean_var+(3*sd_var)
```

The plot of the rank variances (Fig. 3b) can be generated using the plotting code that is provided as a text file (*see* end of Subheading 1.2.3).

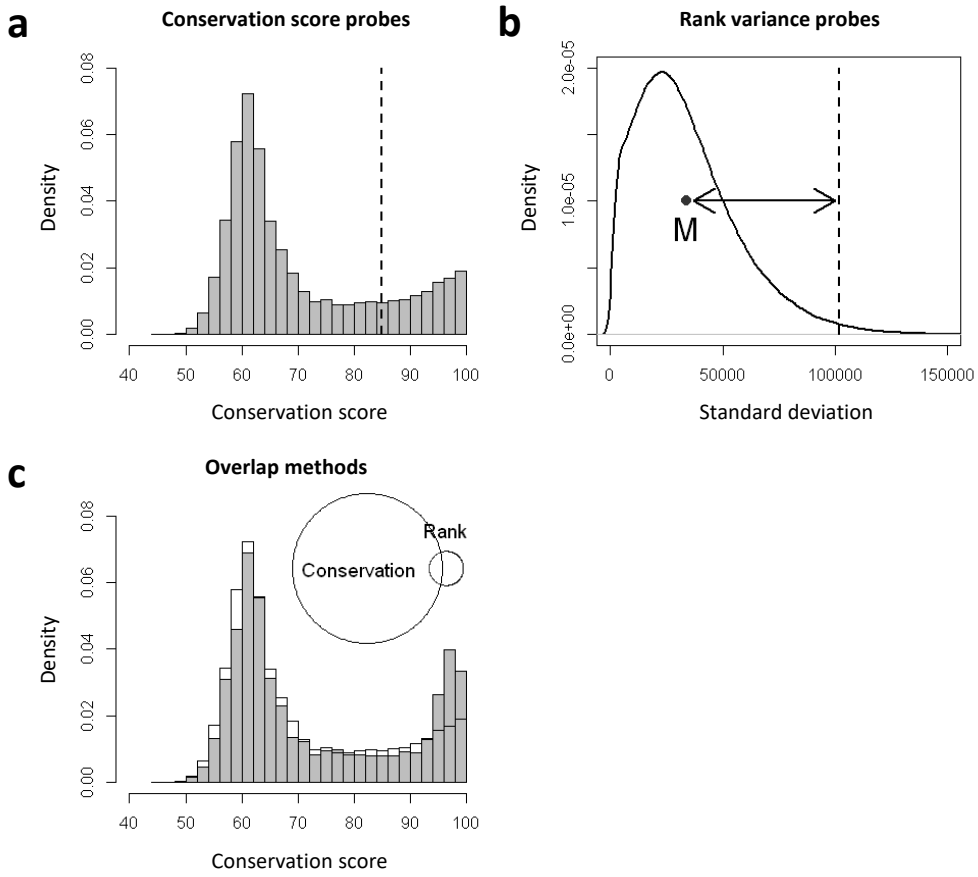

Figure 3. Quality of individual probes. (**a**) Density histogram of the conservation score of the probes. Probes with a conservation score higher than 85 have a high probability to cross-hybridize and are flagged (probes on the right of the dashed line). (**b**) The rank variance distribution of the probes. The rank variance is expressed as a standard deviation. Probes with an abnormal high rank variance are flagged (probes on the right of the dashed line). (**c**) Density histogram of the conservation score of the rank variance probes that were flagged (gray) on top of the conservation score of all probes (transparent). This picture indicates that the probes with a high rank variance also tend to have a high conservation score. The Venn diagram shows however that there is a poor overlap between probes that are flagged with the two methods.

We find that the use of conservation scores and probe rank variance provides a fairly comprehensive assessment of probe quality. That these two criteria are not redundant is reflected in the limited overlap of identified low quality probes (Fig. 3c). We determine this overlap using the following code:

```
> lowq_pr_rank <- rank_var[which(rank_var[,2] > sd_cutoff),1]
> lowq_pr_cons <- cons_score_probes[which(cons_score_probes[,2] >
+ 85),1]
> lowq_probes  <- union(lowq_pr_rank,lowq_pr_cons)
> num_rank    <- length(setdiff(lowq_pr_rank,lowq_pr_cons)) #Only rank
> num_cons    <- length(setdiff(lowq_pr_cons,lowq_pr_rank)) #Only cons
> num_overlap  <- length(intersect(lowq_pr_rank,lowq_pr_cons))
> lowq_rows    <- which(wt_green_red[,1] %in% lowq_probes)
```

Finally, we plot the overlap between the two methods (Fig. 3c) using the plotting code.

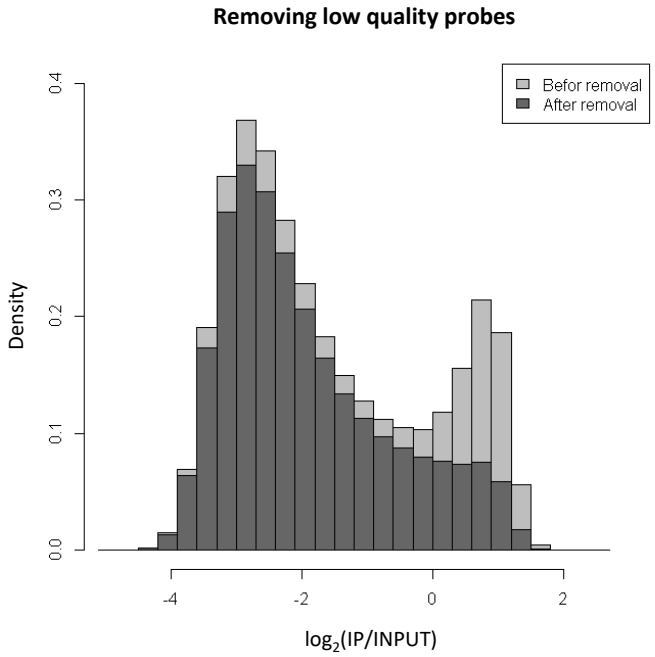

Figure 4. Effect of removing low quality probe signals from the overall $\log_2$(IP/INPUT) signal distribution. Most low quality signals fall in the upper range of the distribution, suggesting that true binding events are partially confounded with cross-hybridization events.

1.3.4.3 The effect of removing low quality probes

The removal of low quality probes has a visible impact on the overall signal distribution. To see this we plot the relative (or ratio) signal of the IP and the INPUT channel in Figure 4 on a $\log_2$ scale (*see* file with plotting code). High signals are typically an indication of increased IP hybridization events relative to the total (INPUT) DNA, thus indexing methylated DNA sequences. We find that most low quality scores fall in the upper signal range, suggesting that true binding events are partially confounded with cross-hybridization events. This is consistent with the observation, in *Arabidopsis*, that DNA methylation primarily occurs in CG-rich repeat elements [4, 5], which have a high cross-hybridization potential. For all subsequent analysis we decided to keep (but flag) low quality probes in the dataset. However, one may also choose to exclude them at this stage.

1.3.5 Implementation of a Hidden Markov Model for reconstructing the DNA methylome

The above-mentioned $\log_2$ transformed IP/INPUT signal ratio is the typical starting point for data analysis. If data from several replicates is available, as in our case, the probe signals can simply be averaged across replicates. We view this distribution (*see* Fig. 4) as a mixture of three partially overlapping components [6]. The right component corresponds to enriched probes (i.e., indexing methylated sequences), the left component to non-enriched probes (i.e., indexing unmethylated sequences) and the middle component to intermediately enriched probes (i.e., indexing intermediately methylated sequences). To illustrate that this mixture view is consistent with the underlying biology, we highlight the probe signals corresponding to annotated transposable elements, which are usually methylated in *Arabidopsis* (Fig. 5a, solid line; [4, 5]). Similarly, as an example of usually unmethylated sequences, we highlight the signal of annotated introns of protein-coding genes (Fig. 5a, dashed line; [4, 5]). The following commands are used to import the probe annotation data. These files simply contain the probe identifiers of probes that match with introns or transposons.

```
> p_id_intron <- read.table(file="intron_probes.txt",
+ header=TRUE,sep="\t")
> p_id_transp <- read.table(file="transposon_probes.txt",
+ header=TRUE,sep="\t")
> head(p_id_intron)
         PROBE_ID
1 CHR01FS000004351
2 CHR01FS000005311
3 CHR01FS000007129
4 CHR01FS000007479
5 CHR01FS000007814
6 CHR01FS000008139
```

For plotting purposes and further analysis steps it is also necessary to know the rows of the probes that correspond to transposons or introns. The following commands determine those rows:

```
> rows_intron <- which(wt_green_red[,1] %in% p_id_intron[,1])
> rows_transp <- which(wt_green_red[,1] %in% p_id_transp[,1])
> rows_intron_highq <- setdiff(rows_intron,lowq_rows) #Without flagged
> rows_transp_highq <- setdiff(rows_transp,lowq_rows) #probes
```

The file with plotting code contains the code for plotting Figure 5. As can be seen in this figure, even within these two extreme annotation sets (i.e., transposons and introns) there is substantial signal variation (Fig. 5). This is probably due to some level of biological variation (i.e., not all transposable element sequences are methylated and not all introns are unmethylated), but it certainly also reflects the limitations of the measurement technology itself [7]. In addition, many probe signals belong to annotation sets that cannot be easily assigned to these extreme mixture components and their classification as methylated, unmethylated, or intermediate is inherently probabilistic.

Our principle analytical approach for performing this probabilistic classification relies on the use of a HMM. A Markov chain is a list of random values $\{H_1, H_2, ..., H_n\}$ that satisfy the so-called Markov property: the value at position i(H_i) is related solely to the values at positions i−1 and i+1 (H_{i-1} and H_{i+1}), with given transition probabilities. In the case of a Hidden Markov chain, an output $\{O_1,O_2, ..., O_n\}$ is observed that depends on the unobserved (hidden) states of the chain, $\{H_1, H_2, ..., H_n\}$ [8]. In the case under consideration, the output or observed chain is the $\log_2$ transformed IP/INPUT signal ratio, while the hidden chain is the methylation state of the DNA sequence indexed by the array probe (Fig. 6).

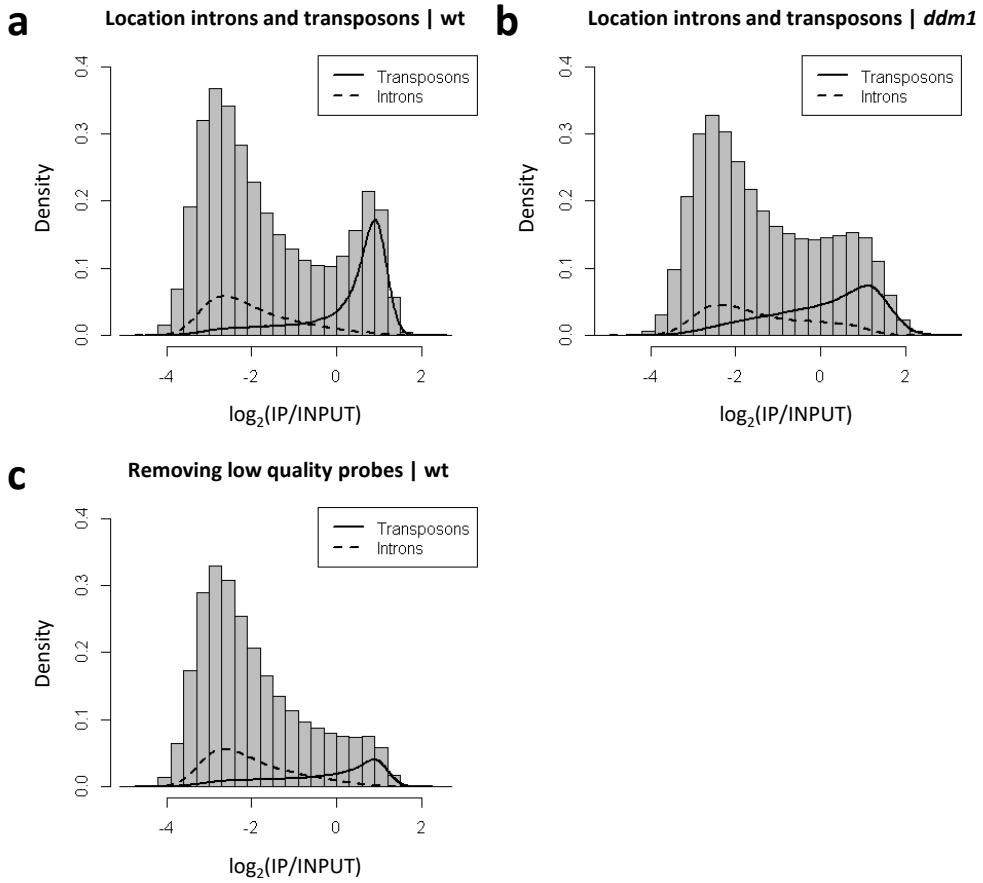

Figure 5. Probe signal distributions of transposable elements and introns. (**a**) The $\log_2$(IP/INPUT) signal distribution of one dye combination (IP: green, INPUT: red) of the wild-type Columbia plant with transposons (solid) and introns (dashed) highlighted. (**b**) Same as in (**a**) but shown for a *ddm1* mutant plant which has lost 70 % of its DNA methylation. The intron distribution is not much affected by this loss. (**c**) Same as in (**a**) but with low quality probe signals removed. As can be seen, the intron distribution is robust to low quality probe signals.

Hence, the HMM approach capitalizes on two key properties of MeDIP-chip data: (1) probe signals are noisy proxies of an unobserved (hidden) methylated, intermediate or unmethylated state, and (2) the probe signals are spatially correlated along the genome, so that neighboring probes provide similar information (Fig. 6). HMMs account for these two properties and provide a powerful statistical framework for

classifying individual probe signals given the overall data structure. Our implementation goal is to provide a robust and fast model estimation procedure. We achieve this by implementing software code in C++ and by incorporating several useful biological constraints. In what follows we outline our version of a HMM that is specifically designed for *Arabidopsis* NimbleGen MeDIP-chip data. We start by detailing key data preparation steps before we move on to discuss the actual implementation strategy.

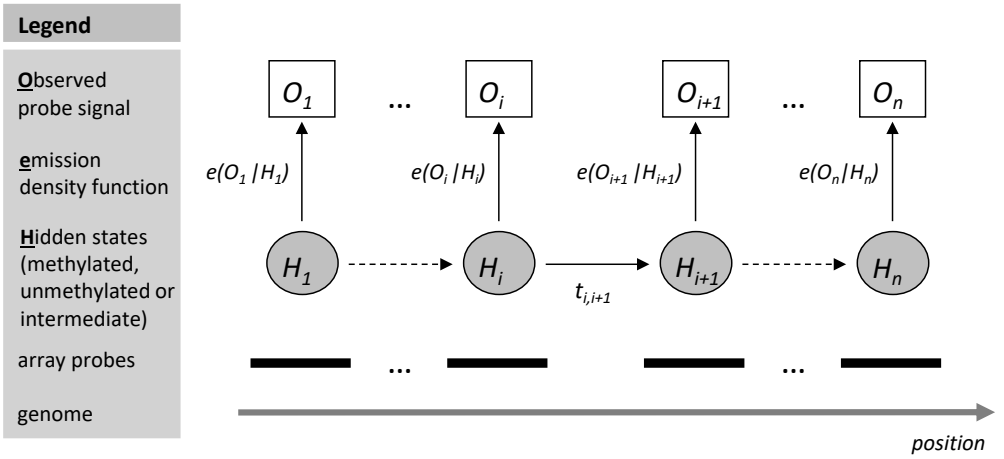

Figure 6. Schematic of a HMM model in the context of genome-wide tiling array data. An explanation of the different components of the HMM is provided in the figure.

1.3.5.1 Data rescaling using intron probes

In the context of a single MeDIP-chip experiment within-array normalization is not required in our experience. Nonetheless, we find that rescaling the overall signal distribution is generally a good idea to permit more meaningful comparisons across different individuals (i.e., experimental conditions), should such additional data become available. To achieve this, we make use of the intron probe signal distribution (Fig. 7a). We standardize this distribution and express the overall signal distribution in terms of their standard deviation values. This has the effect of placing the mean of the intron probe signal at zero and rescaling the values as standard deviation values. This rescaling can be implemented with the following code:

```
> intron_mean    <- mean(wt_dye_swap[rows_intron,2])
> intron_sd      <- sd(wt_dye_swap[rows_intron,2])
> wt_dye_swap_rs <- (wt_dye_swap[,2]-intron_mean)/intron_sd
> wt_dye_swap_rs <- data.frame(wt_dye_swap[,1],wt_dye_swap_rs)
> names(wt_dye_swap_rs) <- c("PROBE_ID","DYE_SWAP_RS")
```

One can use the plotting code that is provided as a text file to plot the density of the rescaled data (Fig. 7a).

We find that the intron signal distribution can be safely used for this rescaling process, insofar that it is relatively invariant to high levels of experimental variation. To illustrate this in the context of an extreme case, we compare the signal distribution for wild-type to that for the *ddm1* mutant, in which DNA methylation is reduced approximately 70 %. The MeDIP-chip experiment reflects this methylation loss nicely (Fig. 5b), with the signal distribution being clearly reduced in height over the enriched component. Clearly, the signal distribution for intronic sequences is not noticeably affected in *ddm1*, as expected.

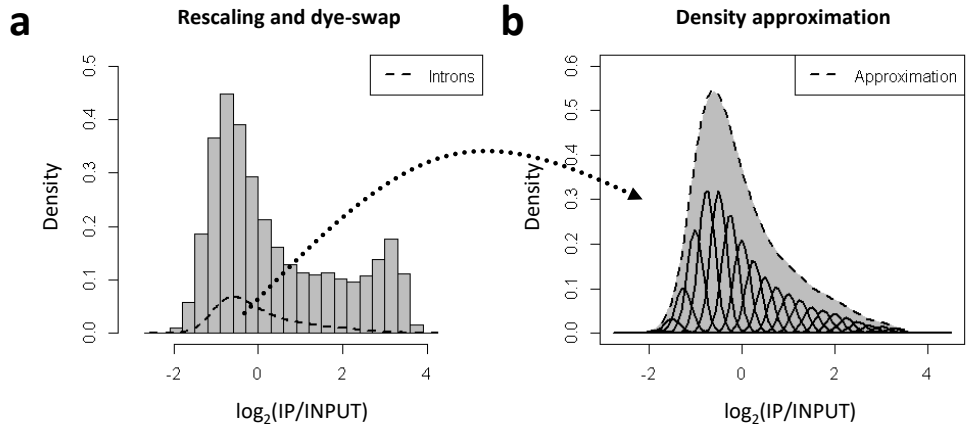

Figure 7. Data rescaling and density approximation probe signal distribution of introns. (**a**) Original signal distribution with intron density highlighted (dashed line). (**b**) Density of the signal distribution for introns approximated using a mixture of a large number of Gaussian distributions with fixed variance and equally spaced means.

1.3.5.2 Implementation of the Hidden Markov Model

We apply our HMM to the rescaled data following a two-step process. First, we use the Baum–Welch algorithm [8, 9] to estimate the best model parameters given the

observed probe signals (Fig. 6). Second, we find the most likely hidden sequence of probe states given these estimated parameters. A copy of the C++ code that we implement can be found at the above URL (*see* end of Subheading 1.2.3). A characteristic feature of our HMM implementation is the use of biologically meaningful constraints on the emission probability density functions, $e(O_i|H_i)$, during the Baum–Welch estimation procedure (Fig. 6). In the following we outline these assumptions. A summary of them can be found in Table 5. Alternatively, all the parameters of the emission probabilities could be freely estimated by means of the Baum–Welch algorithm, but we find that a more biologically meaningful approach is preferable.

Emission probability of unmethylated hidden state: We employ the signal distribution for introns to obtain an approximation of the emission probability of the unmethylated hidden state (Fig. 7a). In this way we incorporate biological knowledge of introns being mostly unmethylated directly into the estimation procedure. This bypasses the need to explicitly assume an emission density function and also speeds up computation. We approximate the signal distribution for introns to an arbitrary degree using mixtures of a large number of Gaussian random variables (Fig. 7b).

Table 5. Summary of the constraints for the emission probability density functions used in the Baum–Welch algorithm.

Hidden state	Distribution	Parameters
Unmethylated states	intron signal distribution	estimated as a mixture of 30 normals (EM algorithm) with fixed variance.
Methylated states	Gaussian	Mean: fixed at the 99th quantile of the intron signal distribution. Variance: freely estimated.
Intermediate states	Gaussian	Mean: fixed at ½ (mean of the methylated distribution). Variance: equal to the variance of the methylated distribution.

Estimation is carried with the EM algorithm [10], which can be implemented using the following code:

```
> density_approx <- function(data,mu,var,lambda,eps,num_norm){
+   mu_diff          <- mu[2]-mu[1]
+   min_val          <- mu[1]-5*mu_diff
+   max_val          <- mu[num_norm]+5*mu_diff
+   rows_extr        <- which(data < min_val | data > max_val)
```

```
+   if(length(rows_extr) > 0){                  # Remove extreme
+     data          <- data[-rows_extr]         # data points
+   }
+   loglik_diff     <- 100000                   # Initial loglik diff
+   counter         <- 0                        # Iteration counter
+   dnorm_tot       <- rep(0,length(data))
+   for(A in 1:num_norm){
+     dnorm_tot     <-
+     dnorm_tot + lambda[A]*dnorm(data,mean=mu[A],sd=sqrt(var))
+   }
+   loglik_pre      <- sum(log(dnorm_tot))      # Initial loglik
+   while(loglik_diff > eps){                   # Estimate mixture
+     counter       <- counter+1
+     for(A in 1:num_norm){                     # Update lambda
+       post        <-                          # Posterior prob
+       lambda[A]*dnorm(data,mean=mu[A],sd=sqrt(var))/dnorm_tot
+       lambda_new  <- sum(post)/length(data)
+       lambda[A]   <- lambda_new
+     }
+     dnorm_tot     <- rep(0,length(data))
+     for(A in 1:num_norm){
+       dnorm_tot   <-
+       dnorm_tot + lambda[A]*dnorm(data,mean=mu[A],sd=sqrt(var))
+     }
+     loglik_new    <- sum(log(dnorm_tot))          # New loglik
+     loglik_diff   <- abs(loglik_new - loglik_pre) # New loglik diff
+     loglik_pre    <- loglik_new
+     cat("Iteration = ",counter," Log-lik diff = ",loglik_diff,"\n")
+    }
+    output         <- list(mu,var,lambda)          # Return results
+    names(output) <- c("mu","var","lambda")
+    return(output)
+ }
>
> mus <- round(seq(-2.75,4.5,0.25),2)
> intron_data <- wt_dye_swap_rs[rows_intron,2]
> den_appr <- density_approx(data=intron_data,mu=mus,var=0.03,
+ lambda=rep((1/30),30),eps=0.1,num_norm=30)
Iteration =  1  Log-lik diff =  64314.34
Iteration =  2  Log-lik diff =  510.154
Iteration =  3  Log-lik diff =  41.84451
Iteration =  4  Log-lik diff =  6.50513
Iteration =  5  Log-lik diff =  1.766472
Iteration =  6  Log-lik diff =  0.7541914
Iteration =  7  Log-lik diff =  0.421424
Iteration =  8  Log-lik diff =  0.2723324
Iteration =  9  Log-lik diff =  0.1922282
Iteration =  10  Log-lik diff =  0.1442565
Iteration =  11  Log-lik diff =  0.1132907
Iteration =  12  Log-lik diff =  0.09209521
>
```

The code for plotting the result (Fig. 7b) is provided as a text file. We generally find that a fit with 30 Gaussians with fixed variance provides a sufficient approximation (Fig. 7b). Parameter estimates are outputted to be used as input in the Baum–Welch algorithm.

Emission probability of methylated hidden state: The second constraint relates to the emission probability for the methylated hidden state. We assume this distribution to be Gaussian, with mean fixed to the 99th quantile of the emission probability of the unmethylated state (i.e., the signal distribution for introns). The variance of the distribution is estimated freely by the Baum–Welch algorithm.

Emission probability of intermediate hidden state: The last constraint relates to the emission probability for the intermediate hidden state. We assume again this distribution to be Gaussian, with a mean that is fixed between the mean of the emission probability of the unmethylated hidden state (i.e., the intron distribution) and the mean of the emission probability of the methylated hidden state. We take the variance of this distribution to be equal to the variance of the emission probability of the methylated hidden state.

The following code generates files that are used as input for the Hidden Markov program written in C++:

```
> values <- c(den_appr$mu,den_appr$var,den_appr$lambda)
> parameters <- c(paste("mu",1:30,sep=""),"var_all",
+ paste("lambda",1:30,sep=""))
> para_est <- data.frame(parameters,values)
> names(para_est) <- c("PARAMETER","VALUE")
> write.table(para_est,"para_wild_type.txt",quote=FALSE,sep="\t",
+ row.names=FALSE,col.names=TRUE)
> write.table(wt_dye_swap_rs,"dye_swap_signal_wild_type.txt",
+ quote=FALSE,sep="\t",row.names=FALSE,col.names=TRUE)
```

Once all the free parameters of the HMM have been estimated, we proceed to infer the most likely hidden sequence of probe states given the parameters of the HMM and the observed probe signals. There are several possible strategies, depending on our optimality criterion. We consider two cases: (1) finding the single best hidden sequence of probe states, given the observed probe signals and the parameters of the HMM (the so-called Viterbi algorithm; [8, 11]), or (2) finding the single hidden probe state which is individually most likely at each position, given the observed probe signals and the parameters of the HMM [8]. A copy of the C++ code implemented for the identification of the optimal sequence according to these two definitions can be found at the above URL (*see* end of Subheading 1.2.3).

1.3.6 Graphical and biological assessment of HMM results

The above algorithms probabilistically classify the original $\log_2$(IP/INPUT) signals to the three underlying methylation states (unmethylated, intermediate or methylated) (Fig. 8). This "hidden chain" of methylation states constitutes the methylome (Fig. 9). Annotation analysis of the probe classification (Fig. 10) shows that most gene probes are unmethylated and the majority of the transposable element probes are methylated, as expected.

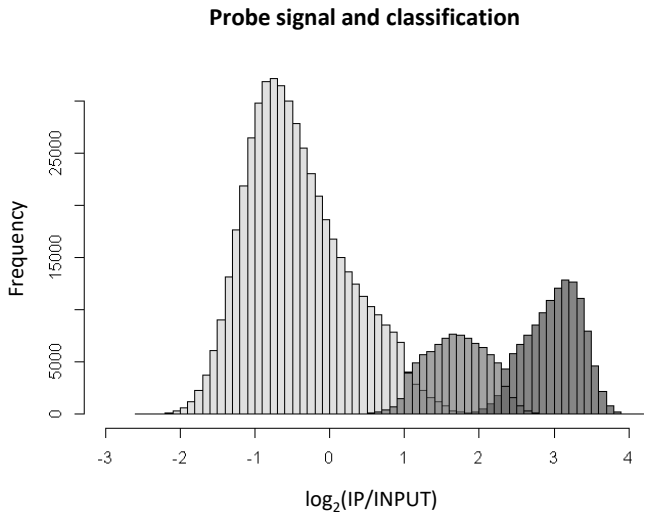

Figure 8. The $\log_2$(IP/INPUT) signal distribution for a wild-type *Arabidopsis thaliana* Col-0 accession. Probes are classified into unmethylated probes (light gray), intermediate probes (gray), and methylated probes (dark gray).

1.4 Conclusions

We have described a comprehensive protocol for the analysis of DNA methylomes in *Arabidopsis* using MeDIP tiling arrays. Our protocol uniquely combines all necessary steps from "wet lab" to "dry lab" to begin to characterize the epigenetic landscape in this species. Owing to the relatively favorable cost of tiling array technology over more recent deep sequencing approaches, our protocol can be easily scaled up to population-level studies. Such large epigenetically informative

approaches will soon become an indispensable tool in the context of intra- or intergenerational functional studies [12]. We have applied the protocol outlined here to a large panel of epiRILs [3] in order to characterize the role of DNA methylation in complex trait inheritance.

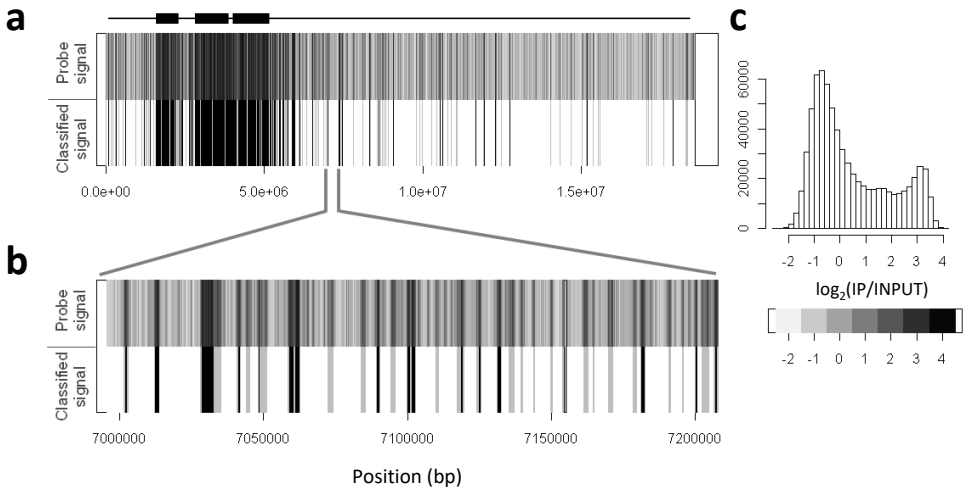

Figure 9. Example probe classification. (**a**) The probe signal (top) and the corresponding (hidden) DNA methylation state (bottom) of chromosome 4 in wild-type Columbia accession plotted against position (base pairs, x-axis); methylated (black), unmethylated (white), and intermediately methylated (gray). As expected, we find substantial methylation in the pericentromeric regions as well as in the heterochromatic knob present on the short arm of the chromosome. (**b**) Magnification of a small region on chromosome 4: we can see how the $\log_2$(IP/INPUT) signal of each probe (top) is assigned to methylated, intermediate, or unmethylated state, depending on its signal and on the signal of its surrounding probes. (**c**) Color code for the probe signal density plot, with the corresponding probe density distribution.

1.5 Notes

1. For more than 250 µL of beads, separate in two tubes for washes.
2. Transfer to new tubes decreases noise. This is done in classical tubes because siliconized tubes tend to leak too much with phenol/chloroform and can cause loss of material.

3. MinElute cleaning is a critical step as the efficiency of WGA2 drops dramatically without it.
4. PicoGreen quantification is very sensitive. Be careful to homogenize your samples well before quantification. Since PicoGreen is not stable in light, quantification must be done soon (less than 30 min) after addition of PicoGreen and samples should be maintained in the dark before use.
5. It is important to verify incorporation of dye using the following formula: concentration in DNA (pmol/µL)/concentration in Dye (pmol/µL). Values are usually between 100 and 180.

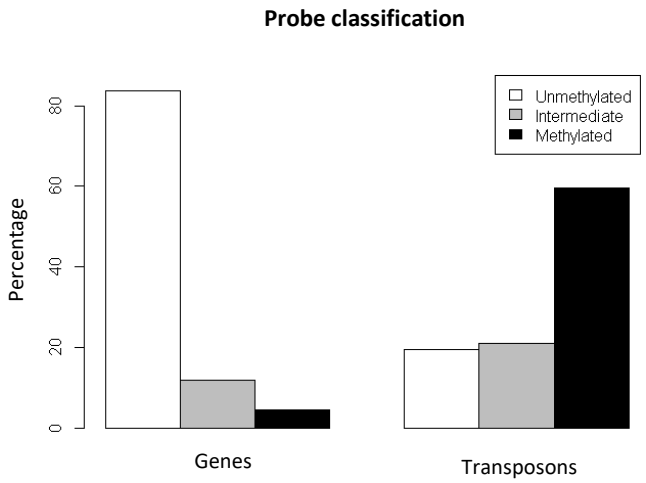

Figure 10. Probe classification of genes and transposable elements.

Acknowledgments

This work was supported in part by grants from the Agence Nationale de la Recherche (Genoplante TAG project, to V.C.) and by the European Union Network of Excellence "The Epigenome" (to V.C.). S.C. was supported by a Ph.D. studentship from the Ministère de l'Enseignement Supérieur et de la Recherche. R.W., M.C.-T. and F.J. were supported by grants from The Netherlands Organization for Scientific Research.

References

1. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, Bouchez D, Dillmann C, Guerche P, Hospital F, Colot V (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5**:e1000530.
2. Reinders J, Wulff BB, Mirouze M, Marí-Ordóñez A, Dapp M, Rozhon W, Bucher E, Theiler G, Paszkowski J (2009) Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes Dev* **23**:939–950.
3. Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E, Labadie K, Wincker P, Jacobsen SE, Jansen RC, Colot V, Johannes F (2012) Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci USA* **109**:16240–16245.
4. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulfite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**:215–219.
5. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**:523–536.
6. Johannes F, Wardenaar R, Colomé-Tatché M, Mousson F, de Graaf P, Mokry M, Guryev V, Timmers HT, Cuppen E, Jansen RC (2010) Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* **26**:1000–1006.
7. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* **11**:191–203.
8. Rabiner LR (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc IEEE* **77**:257–286.
9. Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat* **41**:164–171.
10. McLachlan GJ, Peel D (2000) Finite mixture models. *John Wiley and Sons, Inc.*
11. Viterbi AJ (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inf Theory* **13**:260–269.

12. Johannes F, Colot V, Jansen RC (2008) Epigenome dynamics: A quantitative genetics perspective. *Nat Rev Genet* **9**:883–890.

Chapter 2

Evaluation of MeDIP-chip in the context of whole-genome bisulfite sequencing (WGBS-seq) in *Arabidopsis*

Published as:

Wardenaar R, Liu H, Colot V, Colomé-Tatché M, Johannes F (2013) Evaluation of MeDIP-chip in the context of whole-genome bisulfite sequencing (WGBS-seq) in *Arabidopsis*. *Methods Mol Biol* **1067**:203-224.

Chapter 2

Abstract

Studies of DNA methylation in *Arabidopsis* have rapidly advanced from the analysis of a single reference accession to investigations of large populations. The goal of emerging population studies is to detect differentially methylated regions (DMRs) at the genome-wide scale, and to relate this variation to gene expression and phenotypic diversity. Whole-genome bisulfite sequencing (WGBS-seq) has established itself as a gold standard in DNA methylation analysis due to its high accuracy and single cytosine measurement resolution. However, scaling up the use of this technology for large population studies is currently not only cost prohibitive but also poses nontrivial bioinformatic challenges. If the end-point of the study is to detect DMRs at the level of several hundred base pairs rather than at the level of single cytosines, low-resolution array-based methods, such as MeDIP-chip, may be entirely sufficient. However, the trade-off between measurement accuracy and experimental/analytical practicality needs to be weighted carefully. To help make such experimental choices, we conducted a side-by-side comparison between the popular dual-channel MeDIP-chip NimbleGen technology and Illumina WGBS-seq in two independent *Arabidopsis* lines. Our analysis shows that MeDIP-chip performs reasonably well in detecting DNA methylation at probe-level resolution, yielding a genome-wide combined false-positive and false-negative rate of about 0.21. However, detection can be susceptible to strong signal distortions resulting from a combination of dye bias and the CG content of effectively unmethylated genomic regions. We show that these issues can be easily bypassed by taking appropriate data preparation steps and applying suitable analysis tools. We conclude that MeDIP-chip is a reasonable alternative to WGBS-seq in emerging *Arabidopsis* population epigenetic studies.

2.1 Introduction

DNA methylation is an epigenetic modification that involves the addition of a methyl group to the five position of the cytosine pyrimidine ring. In most animals and plants, this modification has a central role in the regulation of gene expression and in the silencing of transposable elements [1]. Because of its important biological functions, there have been substantial efforts to characterize the complete DNA methylomes of various organisms [2].

In the model plant *Arabidopsis*, the first whole-genome DNA methylation analysis used single-channel Affymetrix and dual-channel NimbleGen tiling arrays [3, 4]. These studies relied on methylation-dependent immunoprecipitation techniques

followed by hybridization to high-density microarray chips (MeDIP-chip), and achieved a resolution of 35 and 220 bp, respectively. This work was instrumental in providing the first picture of the distribution of DNA methylation and its relationship with known sequence annotation in this species. A more detailed view was later obtained by several studies employing Illumina whole-genome bisulfite sequencing (WGBS-seq [5, 6]). WGBS-seq combines bisulfite conversion of DNA with next-generation sequencing (NGS) technologies and provides single cytosine resolution.

While the above-mentioned studies focused on the DNA methylome of a single reference plant, more recent work has begun to document interindividual variation in DNA methylation in large populations. The ultimate aim is to relate this type of epigenetic variation to phenotypic diversity, and to ask broader questions about the role of epigenetics in adaptive evolution. A first step in this direction was recently taken by Schmitz *et al.* [7] and Becker *et al.* [8]. These authors performed WGBS-seq on 8–12 *Arabidopsis* lines and quantified the frequency and distribution of single methylation polymorphisms (SMPs) as well as differentially methylated regions (DMRs). These experiments generated roughly 200–500 GB of data and required construction of extensive data pipelines. Scaling up the use of WGBS-seq to even larger samples poses nontrivial bioinformatic challenges that range from data storage to downstream computation analysis. These challenges can hinder the routine application of this technology for future population epigenetic studies.

A viable alternative is to restrict DNA methylation analysis to the detection of DMRs, which typically range between 10 and 1,000 bp in length. In *Arabidopsis* as in other species DMRs appear to be functionally more important than SMPs [7–10] and appear to be a suitable unit of analysis. Focusing on DMRs has the important advantage that array-based measurement technologies, such as MeDIP-chip, could be employed in place of WGBS-seq because they provide sufficient resolution. The use of array-based methods can substantially reduce the bioinformatic resources required to perform population epigenetic studies. Nonetheless, loss of measurement accuracy resulting from hybridization compared to sequencing may present a significant drawback which not all researchers are willing to accept. Furthermore, it should be noted that unlike WGBS-seq, MeDIP-chip does not allow distinguishing between CG, CHG, and CHH methylation, a point which may be important to consider in some instances. Hence, the trade-off between loss of measurement accuracy and experimental/analytical practicality needs to be weighted carefully. To help make such experimental choices we conducted a side-by-side comparison between the popular dual-channel MeDIP-chip NimbleGen technology and Illumina WGBS-seq. The workflow shown in Figure 1 serves as an outline of this chapter.

Our analysis shows that the dual-channel MeDIP-chip technology performs reasonably well in detection of probe-level DNA methylation, which is approximately the minimum resolution required for DMR detection. We estimate that MeDIP-chip yields a combined false-positive and false-negative rate of 0.21 genome-wide. However, we also find that detection can be critically dependent on prior data preparation steps, signal distortions arising from dye biases, and the statistical method used for detection. Based on our results, we make several simple but important recommendations regarding the experimental implementation of MeDIP-chip for population epigenetic studies (*see* **Note 1–3**).

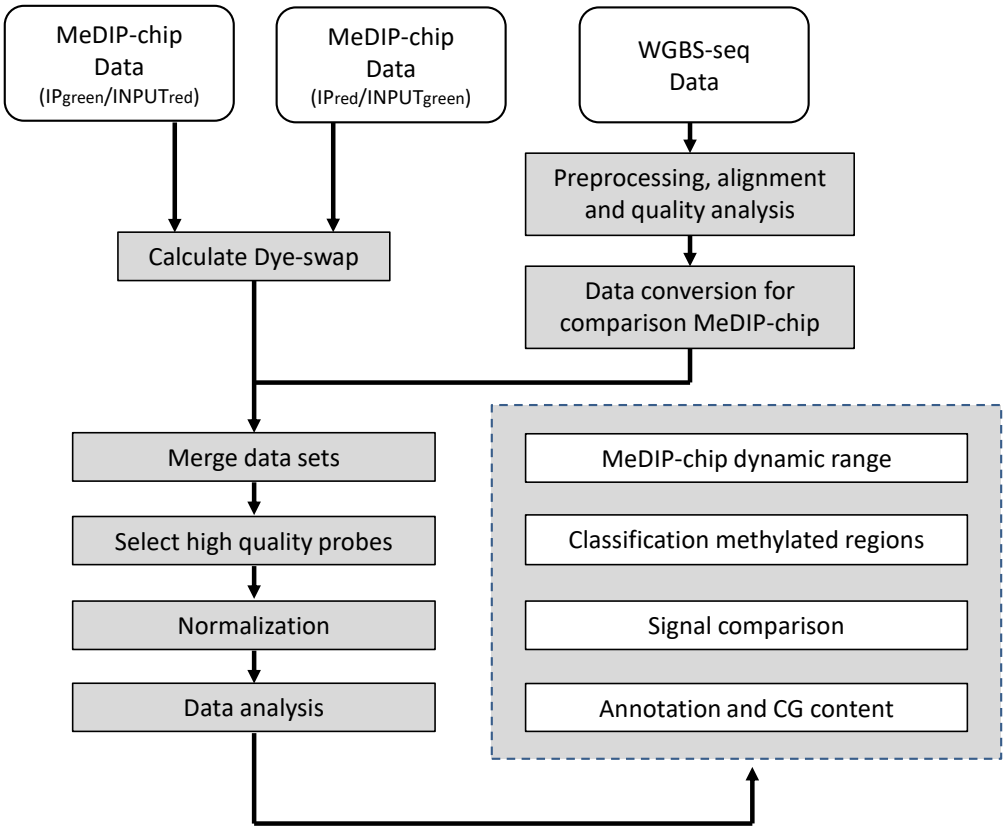

Figure 1. Workflow evaluation MeDIP-chip.

2.2 Data sets and data preparation

In this chapter we consider DNA methylation data from two different epigenetic recombinant inbred lines (epiRILs; R60 and R202 [11]). The DNA methylomes of each epiRIL were measured using dual-channel NimbleGen MeDIP-chip and Illumina WGBS-seq [12]. This section provides a brief overview of these two measurement technologies as well as key data preparation steps.

2.2.1 MeDIP-chip

Methylated DNA immunoprecipitation (MeDIP) is a large-scale purification technique used for the enrichment of methylated DNA fragments. This technique uses antibodies specific to methylated cytosines in order to separate methylated DNA fragments from unmethylated DNA fragments. Following this separation procedure the methylated fragments can either be hybridized to a tilling array (MeDIP-chip [13]) or sequenced (MeDIP-seq [14]) in order to assess the methylation status of the genome under consideration. The application of MeDIP-chip to the two epiRILs under consideration involved several experimental steps which are discussed in short in this paragraph. A more extensive description of the protocol used to obtain MeDIP-chip data described in this chapter is given by Cortijo *et al.* [15].

1. *Extraction and fragmentation* — DNA was extracted from aerial parts of three-week-old *Arabidopsis* plants using a standard extraction kit (Qiagen DNeasy plant Maxi kit). Extracted DNA was fragmented using sonication. Sonication produced fragments with a size between 100 and 600 bp (verified with gel electrophoresis).
2. *Immunoprecipitation and amplification* — After this fragmentation step, the DNA was denatured and anti-5mC antibody was added to the IP DNA pool, which recognizes specifically methylated cytosines in single-stranded DNA. Magnetic beads containing binding sites for this antibody were then added to pull down (i.e., immunoprecipitate) methylated fragments. Following release of the antibody, both IP and input DNA fractions were amplified by PCR.
3. *Labeling and hybridization* — The IP and input fractions were differentially labeled with two fluorescent dyes (Cy3 and Cy5) and hybridized to NimbleGen whole-genome *Arabidopsis* tiling arrays (3×720K array) containing 711,320 isothermal probes. Depending on the CG content, these

probes range from 50 to 75 nucleotides in length and have an inter-probe spacing of about 110 base pairs on average.

4. *Scanning the tiling array* — After hybridization, the intensities of both dyes were obtained by scanning the tiling array. The scanner outputs two files with the raw intensities for each probe on the tiling array: one file with the IP signals and the other file with the input signals.
5. *Signal calculation* — The IP and input signals were log transformed and subtracted from each other ($\log_2$ (IP) − $\log_2$ (input)). Hence, probes with a low signal are from genomic regions that show low methylation levels, and those with a high signal are from genomic regions that show high methylation levels.

2.2.2 Whole-genome bisulfite sequencing

WGBS-seq is a NGS technology used to determine the DNA methylation status of single cytosines. In the case of WGBS-seq, unlike other NGS technologies, the DNA is treated with sodium bisulfite before sequencing. Sodium bisulfite is a chemical compound that converts unmethylated cytosines into uracil [16, 17]. Knowing which cytosines have converted it is possible to determine which cytosines are methylated (not converted) and which ones are unmethylated (converted into U). After sequencing, the unmethylated cytosines appear as thymines. There are several ways of producing sequence libraries for WGBS-seq (*see* Ref. 18). The WGBS-seq data in this chapter was produced with a technique developed by Cokus *et al.* [6]. Here we describe in short the procedure.

1. *Extraction and fragmentation* — DNA was extracted from aerial parts of three-week-old *Arabidopsis* plants using standard extraction kit (Qiagen DNeasy plant Maxi kit). Extracted DNA was fragmented by sonication.
2. *Adapter ligation and size selection* — A set of double-stranded adapter sequences was ligated to the fragmented DNA. These adapter sequences contained methylated adenine bases with DpnI restriction sites. The restriction sites are important for the removal of the first set of adapter sequences in one of the subsequent steps. Gel electrophoresis was used to obtain adapter-ligated fragments with an appropriate size.
3. *Bisulfite conversion and amplification* — After the addition of the adapter sequences the sodium bisulfite conversion is performed. During this step, unmethylated cytosines are changed into uracil. PCR was subsequently

performed with the use of primers that were complementary to the converted adapter sequences.

4. *Removal of first adapter sequences and ligation sequencing adapters* — After the first PCR amplification step the first set of adapters was removed using DpnI restriction enzymes. A new set of sequencing adapters was subsequently ligated to BS-converted DNA fragments.
5. *Size selection and amplification* — Fragments with a size between 120 and 170 bp were selected with the use of gel electrophoresis, and a second and final PCR step was performed using primers complementary to the sequencing adapters to yield a sequencing library.
6. *Sequencing* — Illumina sequencing technology (Illumina 1G Genome Analyzer) was used to produce read sequences with a length of 76 or 101 nt.

After sequencing, the reads need to be mapped (or aligned) to a reference genome in order to infer the methylation status of the cytosines of the genome under consideration. However, mapping these converted sequences is not straightforward since unmethylated cytosines will result in mismatches with the reference genome (i.e., they appear as thymines). To circumvent this issue several programs have been developed that first convert the reference genome into a three-letter genome (i.e., all cytosines are changed into thymines; in silico [19]). The remaining cytosines of the read sequences also need to be changed into thymines before mapping. The in silico-treated read sequences are subsequently mapped to this three-letter reference genome. Once the mapping has been performed, methylation status can be inferred using the original sequence of the reference genome and the read. A thymine (read) mapped to a cytosine (reference) is an unmethylated cytosine. A cytosine mapped to a cytosine is a methylated cytosine. We utilized BS Seeker [20] for the mapping of the read sequences. BS Seeker is a python-based open-source mapping program for the alignment of bisulfite-treated sequences. The analysis includes preprocessing of the reads prior to alignment, the alignment itself, and quality analysis of the data. The steps are described further below. More details of the mapping of the reads can be found elsewhere [12].

7. *Removal of adapter parts* — When a DNA fragment is shorter than the read sequence a part of the adapter sequence will also be sequenced. The adapter sequence was added artificially and does not match with the reference genome. We therefore removed this part using a sliding window

approach. The part that overlapped with the known adapter sequence was removed.

8. *Removal of short reads* — The removal of the adapter sequences resulted in some cases in reads with a length smaller than 30 nucleotides. These short reads are more difficult to map and were therefore removed. The proportion of short reads was in our case quite small and therefore the final read coverage was barely affected.
9. *Removal of duplicated reads* — Duplicated reads were removed in the final preprocessing step because they were likely produced during PCR amplification and were therefore not informative.
10. *Alignment to reference genome* — After these preprocessing steps the reads were mapped to a reference genome (TAIR 10) with the use of BS Seeker. After mapping the obtained average genome coverage was 29 and 27× for epiRIL 60 and 202, respectively (both strands combined).
11. *Determination conversion rate* — One important step in the quality analysis of the data is to determine the (bisulfite) conversion rate. The conversion rate, which is the percentage of unmethylated cytosines that effectively changed into uracil, was determined for both epiRILs after mapping. The conversion rates were determined with the information of reads that were mapped to chloroplast DNA. The chloroplast DNA is known to be unmethylated and therefore any detected methylated cytosine is considered to be a non-converted unmethylated cytosine. Both epiRILs showed a conversion rate above 99 % which indicates that the data is of good quality.

2.2.3 Data conversion and normalization

One of the major differences between MeDIP-chip and WGBS-seq is mapping resolution. WGBS-seq can interrogate the methylation status of individual cytosines while MeDIP-chip achieves a resolution of about 165 bps. In order to facilitate a meaningful comparison between the two technologies, we converted the WGBS-seq data into a format that is comparable to that of MeDIP-chip. To achieve this we calculated the proportion of methylation calls in windows of 165 bps centered at the probe sequence (Fig. 2). By methylation calls we mean the individual methylation calls of each read sequence. This results in a signal ranging from zero to one. This signal is afterwards normalized for the number of cytosines in the probe window. Let C_j^m denote the number of cytosines that have been called methylated in the j^{th} window, C_j^u the number of cytosines that have been called unmethylated in the j^{th}

window, R_j the total number of cytosines in the jth window according to the reference genome, and R_{max} the maximum number of cytosines across all windows. The converted and normalized WGBS-seq signal can be calculated as

$$WGBS_{sig}(j) = \frac{C_j^m}{C_j^m + C_j^u} \times \frac{R_j}{R_{max}}$$

We selected high-quality data for the analysis described in this chapter. In case of the WGBS-seq probe windows, we only selected windows with 35 or more cytosines, with at least half of the cytosines being covered by one or more read sequences. In case of the MeDIP-chip data we only selected probes with a conservation score smaller or equal to 85. The conservation score of a probe indicates the uniqueness of a probe sequence. These scores were obtained by performing a blast search. Scores are percentage of identity with the second best hit (score range 45–100). The best hit is with the genomic location for which the probe was designed. Probes with a high conservation score are more likely to cause cross-hybridization problems. As shown in Figure 3, removal of probes with a high conservation score has a drastic impact on the MeDIP signal distribution, as they typically show signal intensities similar to probes that correspond to genomic regions with high methylation levels. *See* **Note 1** for recommendations concerning the quality of the data.

Analysis was performed on probes (i.e., probe windows) that were present in both data sets (MeDIP-chip and WGBS-seq signal data). This yielded 551,688 and 550,676 high-quality probe windows in total, covering approximately 77.6 and 77.4 % of the genomes of R60 and R202, respectively. In case of the MeDIP-chip data a total of two dye-swap experiments were performed for each epiRIL. The log-transformed signals were also averaged over the two dye-swap experiments which resulted into three MeDIP-chip data sets:

- G/R data: IP labeled green (Cy3, G) and input labeled red (Cy5, R).
- R/G data: IP labeled red (Cy5, R) and input labeled green (Cy3, G).
- DS data: Average of G/R and R/G data (dye-swap data).

Finally, quantile normalization was applied to bring these three data sets to a common scale [21]. Figure 4 displays a density histogram of the WGBS-seq signal for the two epiRIL experiments (R60 and R202) and the MeDIP-chip signal (DS).

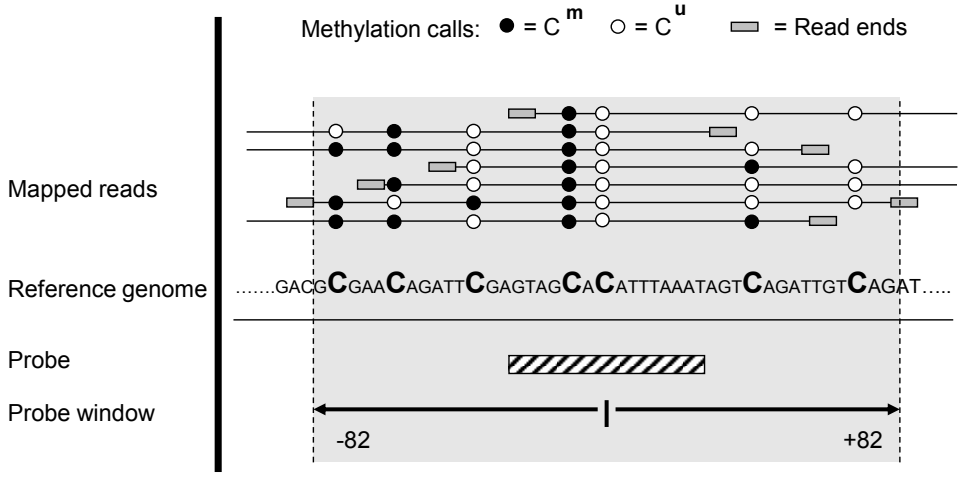

Figure 2. Calculation of whole-genome bisulfite sequencing signals. The methylation calls within the window (gray) are used to calculate a normalized whole-genome bisulfite sequencing signal (*see* formula).

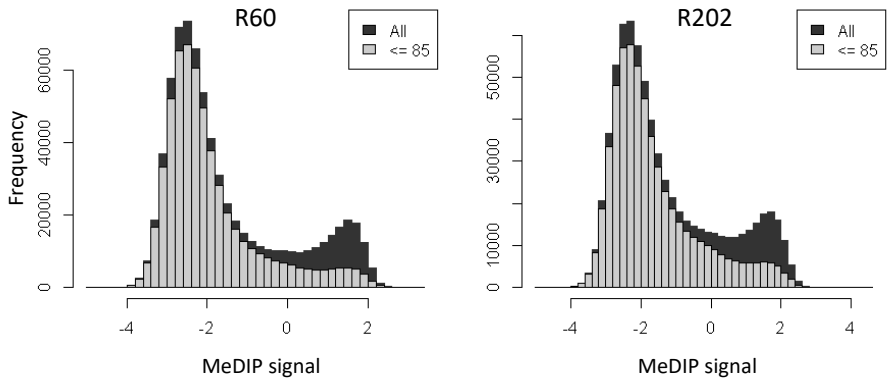

Figure 3. Impact of removing probes with a high conservation score on probe signal distribution. Probes with a high conservation score (dark gray) typically show signal intensities similar to probes that correspond to genomic regions with high methylation levels.

2.2.4 Software

The analysis described in this chapter was performed using R [22]. R is a command-line software environment for statistical computing and graphics. The Hidden Markov Model (HMM) for probe classification was programmed in C++ [15].

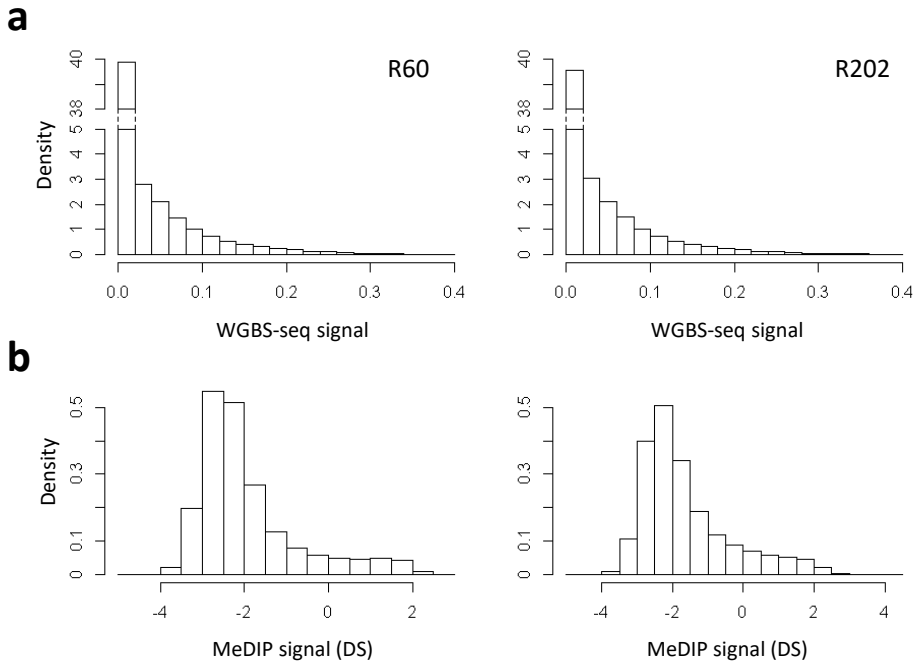

Figure 4. Density histogram of WGBS-seq signals and MeDIP-chip signals following data conversion and normalization. (**a**) WGBS-seq signal distribution. (**b**) MeDIP-chip signal distribution.

2.3 Results

2.3.1 Assessment of MeDIP-chip dynamic range

Since the WGBS-seq signal provides a measure of the proportion of methylated cytosines in a given probe window, we were able to assess the dynamic range of the MeDIP-chip technology directly by empirical comparison. To achieve this we

calculated the median MeDIP-chip signal for sliding windows along the entire WGBS-seq signal range (Window size: 0.05; step size: 0.01; Fig. 5a). We find that MeDIP-chip exhibits good sensitivity for low-to-intermediate methylation levels (WGBS-seq range: 0.00 to ~0.13). In this low-to-intermediate range, there is a nearly linear relationship between the WGBS-seq signal and the MeDIP median signal, but the MeDIP-chip sensitivity falls off quickly and saturates at a WGBS-seq signal value of about 0.28. Above this point, MeDIP-chip is effectively unable to differentiate between methylation levels. However, in R60 and R202 this saturation effect affects only a relatively small number of probe windows, 0.19 % (*N* = 1,056) and 0.22 % (*N* = 1,220) of all regions genome-wide, and 0.58 and 0.67 % of all methylated regions (*see* Table 1), respectively. Signal saturation should therefore not be a matter of great concern in the analysis of the *Arabidopsis* methylome. Similar conclusions can be reached when considering the MeDIP signal on its original scale (data not shown),

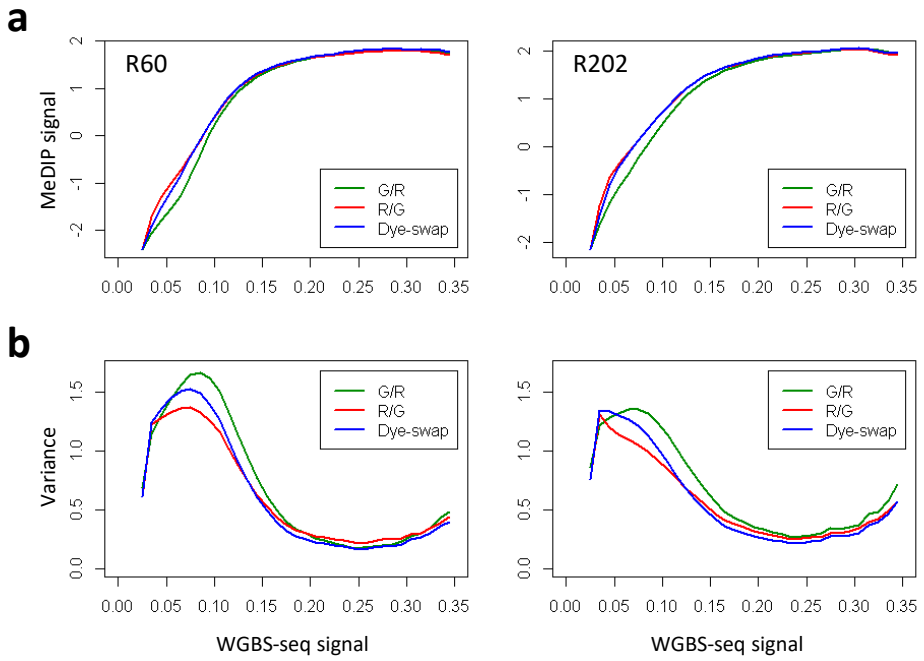

Figure 5. Median and variance of the MeDIP signal along the entire WGBS-seq signal range. Median MeDIP signal (**a**) and variance MeDIP signal (**b**) for sliding window along the entire WGBS-seq signal range. The G/R data show less sensitivity for low-to-intermediate WGBS-seq signals and also show a higher variance compared to the R/G data.

rather than on the log-transformed scale, which indicates that saturation is not caused by scaling issues.

Our analysis also indicates that there are clear dye-related differences in the MeDIP median signal. The R/G data appears to respond more sensitively to changes in WGBS-seq methylation levels compared with the G/R and the DS data (Fig. 5a). While these dye differences disappear at the saturation point (WGBS-seq signal ~0.28), they are most prominent in the optimal dynamic range. This divergence is even more severe when we consider the MeDIP signal variance across the complete WGBS-seq range using the same sliding window approach (Fig. 5b). Ideally, the signal variance should be low and constant across methylation levels. Figure 5b illustrates that this is clearly not the case: the MeDIP signal variance is largest within the optimal dynamic range but decreases rapidly with increasing methylation levels. Notably, the G/R data displays a 1.20 (R60)- and a 1.28 (R202)-fold increase in signal variance (on average) relative to the R/G data suggesting that it is substantially noisier, and the dye-swap (DS) fails to correct this bias. This latter observation is contrary to what is typically seen in expression microarrays where dye-swaps have proved to be an effective strategy [23, 24]. *See* **Note 2** for recommendations concerning the labeling of the IP and input DNA.

Table 1. Classification probe windows using WGBS-seq.

	R60	**R202**
Classification cutoff	5.09E-03	7.03E-03
# probe windows selected for analysis	551,688 (77.6)	550,676 (77.4)
# unmethylated probe windows	369,358 (67.0)	367,526 (66.7)
# methylated probe windows	182,330 (33.0)	183,150 (33.3)
Median signal, unmethylated windows	9.58E-4 (0.000 – 5.09E-3)	9.65E-4 (0.000 – 7.02E-3)
Median signal, methylated windows	3.32E-2 (5.09E-3 – 0.552)	3.33E-2 (7.03E-3 – 0.562)

Reported are the number of unmethylated and methylated windows using a genome-wide false-positive rate of 0.01.

2.3.2 Classification of methylated regions using MeDIP-chip

From a data analysis standpoint, the classification of individual probes according to their underlying methylation status (i.e., methylated or unmethylated) is critically

dependent on the variance of the MeDIP signal. When signal variation is large, classification tends to be more difficult. Many statistical analysis methods have been proposed to minimize this problem and to facilitate accurate probe classification in the context of MeDIP-chip data (e.g., [15, 25–29]). The development of such methods continues to be an active area of research. Classification is particularly problematic in regions of the MeDIP signal distribution where signals from unmethylated probes overlap with those from methylated probes (Fig. 6). In this case, the task is to find an informative cutoff that would minimize both false-positive and false-negative methylation calls.

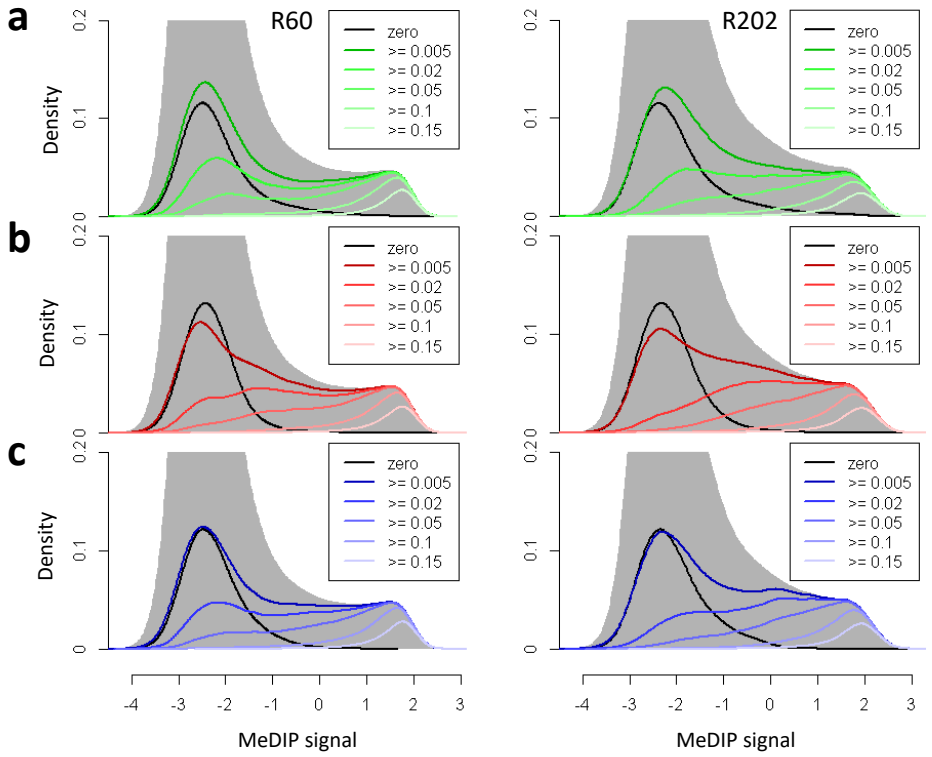

Figure 6. Overlap unmethylated and methylated probe signal distributions. Shown are the MeDIP distributions that correspond to a certain WGBS-seq signal range. It becomes clear that there is a significant overlap between the MeDIP signal distributions with a low WGBS-seq signal (unmethylated; black is WGBS-seq signal of zero) and those with a high WGBS-seq signal (methylated). The G/R data (**a**) tends to have a higher overlap compared to the R/G data (**b**). The DS data (dye-swap; **c**) shows an average overlap result. It also shows a higher overlap compared to the R/G data.

2.3.2.1 Defining the “Gold Standard"

To illustrate this problem empirically we use the WGBS-seq signal to define high-confidence methylated probe windows based on a measurement error distribution. To achieve this, let the measurement error, y, of the j^{th} probe window be given by

$$y(j) = (1 - CR) \times \frac{R_j}{R_{max}}$$

where R_j is the number of cytosines in the j^{th} probe window, R_{max} the maximum number of cytosines across all windows, and CR is the overall conversion rate. Furthermore, let us define the empirical density distribution of the error as $f(y)$ (Fig. 7). Using this distribution, the genome-wide false-positive ($GWFP$) rate can be calculated numerically using

$$GWFP = 1 - \int_0^T f(y)dy$$

where T is the WGBS-seq signal threshold that is needed to meet a given $GWFP$ level. We find that for $GWFP$ = 0.01, the WGBS-seq threshold is approximately 5.09E−3 and 7.03E−3 for R60 and R202, respectively. Hence, we define probe

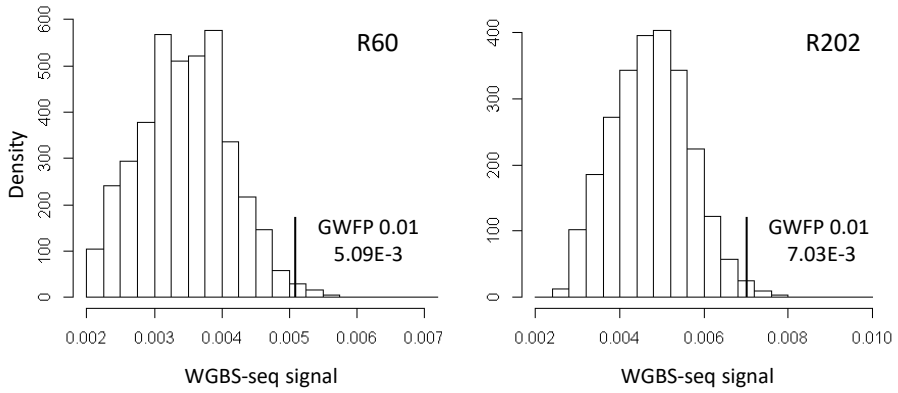

Figure 7. Measurement error distributions. Shown are the measurement error distributions of both epiRILs and the signal cutoffs that correspond to a genome-wide false-positive rate of 0.01.

window j as methylated if the WGBS-seq signal of that region is larger than the threshold T. At this threshold level, we find that 33.0 % (*N* = 182,330) and 33.3 % (*N* = 183,150) of all probe windows (genome-wide) can be confidently called methylated with this technology in R60 and R220, respectively (Table 1).

2.3.2.2 MeDIP signal classification based on a naïve classifier

We use the WGBS-seq-derived classification to assess the problem of determining the methylation status of probes in the context of MeDIP-chip data. We first consider a naïve classifier which consists of a single MeDIP cutoff. According to this classifier, a probe is considered methylated if its signal exceeds the cutoff and as unmethylated if its signal falls below it. Comparing the resulting calls to those obtained from the WGBS-seq classification (Subheading 2.3.2.1) allows us to define the MeDIP false-positive and false-negative rates associated with the naïve classifier. Figure 8a, b shows the distribution of false-positive and false-negative rates for series of cutoffs across the entire MeDIP signal range (−4 till 3 with step size 0.2). This analysis shows that there is a considerable trade-off between minimizing false positives and false negatives while maximizing the total number of regions detected as methylated. We find that the optimal cutoff corresponds to a MeDIP signal of about –1.5 (Fig. 8), which yields a combined false-positive and false-negative rate of about 0.23. Consistent with the dye-effects illustrated in our assessment of the dynamic range (Subheading 2.3.1), the combined false-positive and false-negative rates show

Table 2. Number of false-positive and false-negative probe classifications.

	R60			R202		
	FP	FN	FP + FN	FP	FN	FP + FN
G/R – naïve cutoff	17,973	120,335	138,308 (0.251)	25,945	111,320	137,265 (0.249)
G/R – HMM	12,796	114,197	126,993 (0.230)	12,385	110,766	123,151 (0.224)
R/G – naïve cutoff	16,763	97,826	114,589 (0.208)	24,932	83,325	108,257 (0.197)
R/G – HMM	11,564	108,238	119,802 (0.217)	13,985	101,413	115,398 (0.210)
DS – naïve cutoff	14,901	107,635	122,536 (0.222)	15,852	101,227	117,079 (0.213)
DS – HMM	12,905	107,769	120,674 (0.219)	15,179	100,496	115,675 (0.210)

Reported are the number of false-positive and false-negative probe classifications using either a naïve cutoff or a Hidden Markov Model. Numbers of the naïve cutoff are based on the most optimal cutoff in Figure 8c (smallest number of FP + FN).

substantial dye-dependence, with genome-wide rates of about 0.25, 0.20, and 0.22 for the G/R, R/G, and the DS data, respectively (Fig. 8; Table 2). Hence, the G/R data

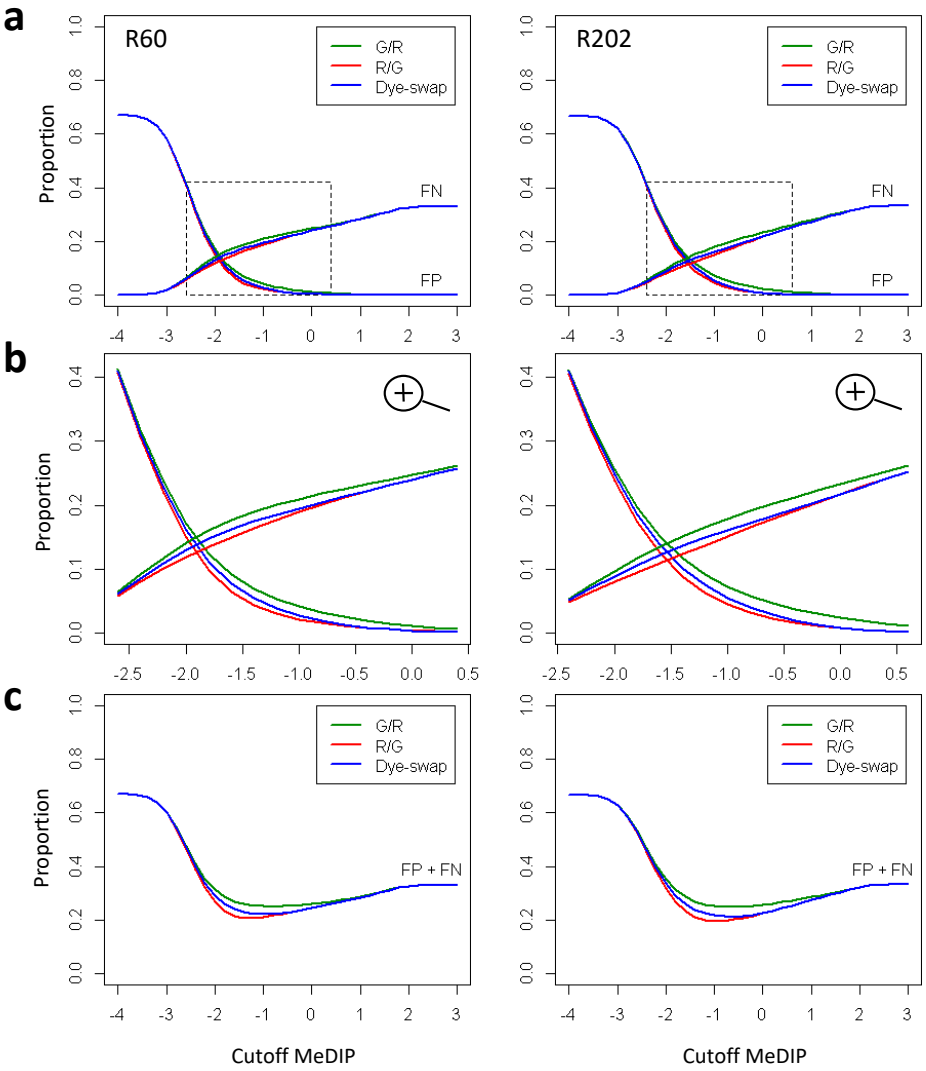

Figure 8. False-positive and false-negative rates using a naïve classifier. Shown are the proportions of false-positive and false-negative probe classifications for different classification cutoffs of the MeDIP data (**a**) and (**b**) and the sum of the two (**c**). The G/R data shows a substantial higher proportion of FP and FN compared to the R/G data.

yields the highest misclassification rate which is likely caused by its high signal variance. Again, the dye-swap does not correct this problem (Fig. 8). *See* **Note 2** for recommendations concerning the labeling of the IP and input DNA.

2.3.2.3 MeDIP signal classification based on Hidden Markov Model

This dye bias can be partially alleviated if one considers, instead of the above naïve classifier, a more sophisticated statistical classification approach based on HMMs (Fig. 9, Table 2). We have recently proposed an HMM for the analysis of MeDIP-chip [15]. This model has been shown to outperform alternative methods in terms of speed, sensitivity, and specificity [28]. An important aspect of this model is, as with HMMs in general, that it borrows signal information from immediately surrounding probes, and therefore significantly reduces measurement noise. This leads to a more robust inference of the underlying methylation status of a given probe window and makes methylation analysis less susceptible to dye effects. Figure 9 illustrates this point clearly; it shows that the HMM analysis of the R/G, G/R, and DS data results in much smaller misclassification differences between these data sets in terms of overall false-positive and false-negative rates (about 0.23, 0.21, and 0.21 for the G/R, R/G, and the DS data, respectively; Table 2). *See* **Note 3** for recommendations concerning the analysis of the data. Nonetheless, despite this improvement, dye-related differences, particularly in the G/R data, do persist and continue to affect our ability to infer the correct methylation status of a given genomic region (Fig. 9). It is

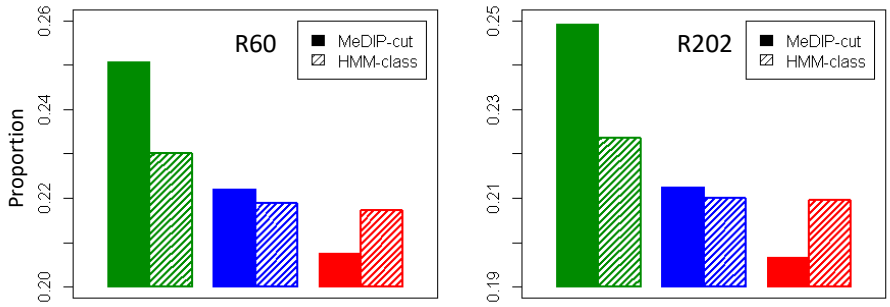

Figure 9. Comparison performance of a naïve classifier and a Hidden Markov Model. Shown are the proportion of misclassified probes (FP + FN) obtained using the most optimal MeDIP classification cutoff (MeDIP-cut; smallest number in Fig. 8c) or the HMM classification (HMM-class). The color of each bar corresponds to the three data sets (green: G/R data; red: R/G data; blue: DS data).

therefore of interest to identify and characterize the sources of this bias in the MeDIP-chip data.

2.3.3 Dye bias in MeDIP-chip is associated with low methylation levels and CG content

To explore the source of the observed dye bias, we start by plotting the two dye combinations (G/R and R/G) in a scatter plot (Fig. 10a). We find that there is a subset of the probes that shows a relatively higher signal for the G/R data compared to the R/G data. Inspection of the WGBS-seq signal corresponding to these probes indicates that the methylation level of these probe windows should be low (high density of probes around zero). This expectation is indeed reflected in the R/G signal (Fig. 10c), but not in the G/R signal which seems to be vastly exaggerated (Fig. 10b). Since the dye-swap signal yields only an average of the R/G and G/R data it cannot correct this bias (Fig. 10d).

Annotation analysis of this subset of probes shows that they contain a high proportion of transposon and intergenic sequences relative to genic sequences (Fig. 11a). In *Arabidopsis*, it is well known that genes have a higher CG percentage compared to transposons (Fig. 12). This raises the question whether CG content may be a key contributor to the observed dye bias. In order to explore this possibility more generally, we calculated the CG content of the probe window for each probe on the tiling array and examined its relationship with signal intensity in the G/R, R/G, and DS data sets. For clarity we restricted our analysis to probes that were unmethylated according to WGBS-seq (*see* Table 1). In this way we could rule out any trends arising from differences in methylation levels. Our analysis shows that the signal intensity of unmethylated regions in the G/R data is subject to strong dye biases (Fig. 11b). We find a clear negative linear relationship between CG content and signal intensity; that is, signal intensity is highest for probes with low CG content and lowest for probes with high CG content. By contrast, CG content appears to have little influence on the signal intensity in the R/G data (Fig. 11c), and the DS displays intermediate levels of CG bias (Fig. 11d). *See* **Note 2** for recommendations concerning the labeling of the IP and input DNA.

Royce *et al.* [30] considered normalizing tiling array signals for the CG content of probes. While this procedure may work for some applications, such as transcription factor binding data (ChIP-chip), its application to gene expression tiling arrays has been shown to lead to overnormalization and hence to a loss of signal information [31]. Overnormalization is expected to be even more drastic in MeDIP-

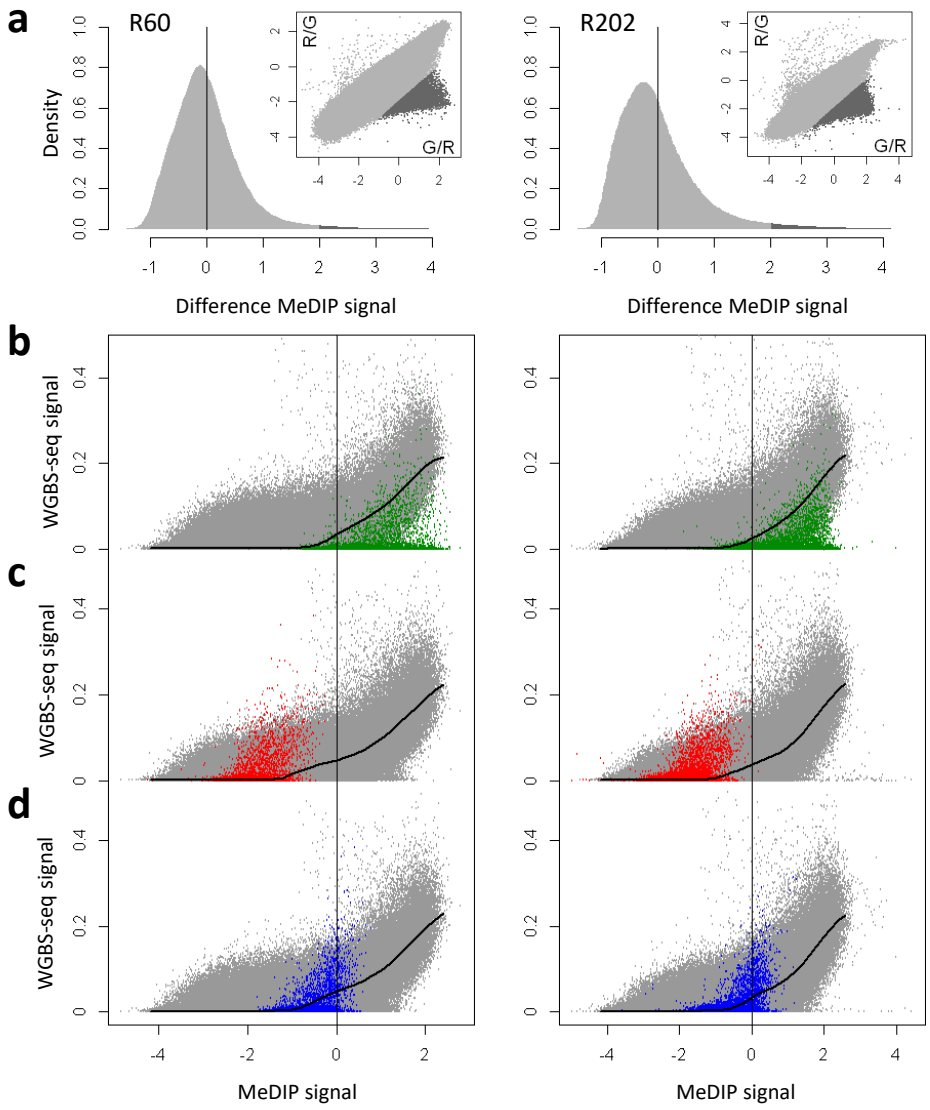

Figure 10. Inspection of non-correlating probes using WGBS-seq signal data. (**a**) Shown is a density plot of the difference of both dye combinations. The inset shows a scatter plot of both dye combinations. Probes with a signal difference higher than two are indicated with dark gray. (**b**)–(**d**) Scatter plots of WGBS-seq data (y-axis) and each of the three MeDIP data sets (x-axis). The non-correlating probes (dark gray in panel **a**) are highlighted according to the color assigned to each data set (green: G/R data; red: R/G data; blue: DS data). The black line shows the median WGBS-seq signal for sliding windows along the entire MeDIP signal range.

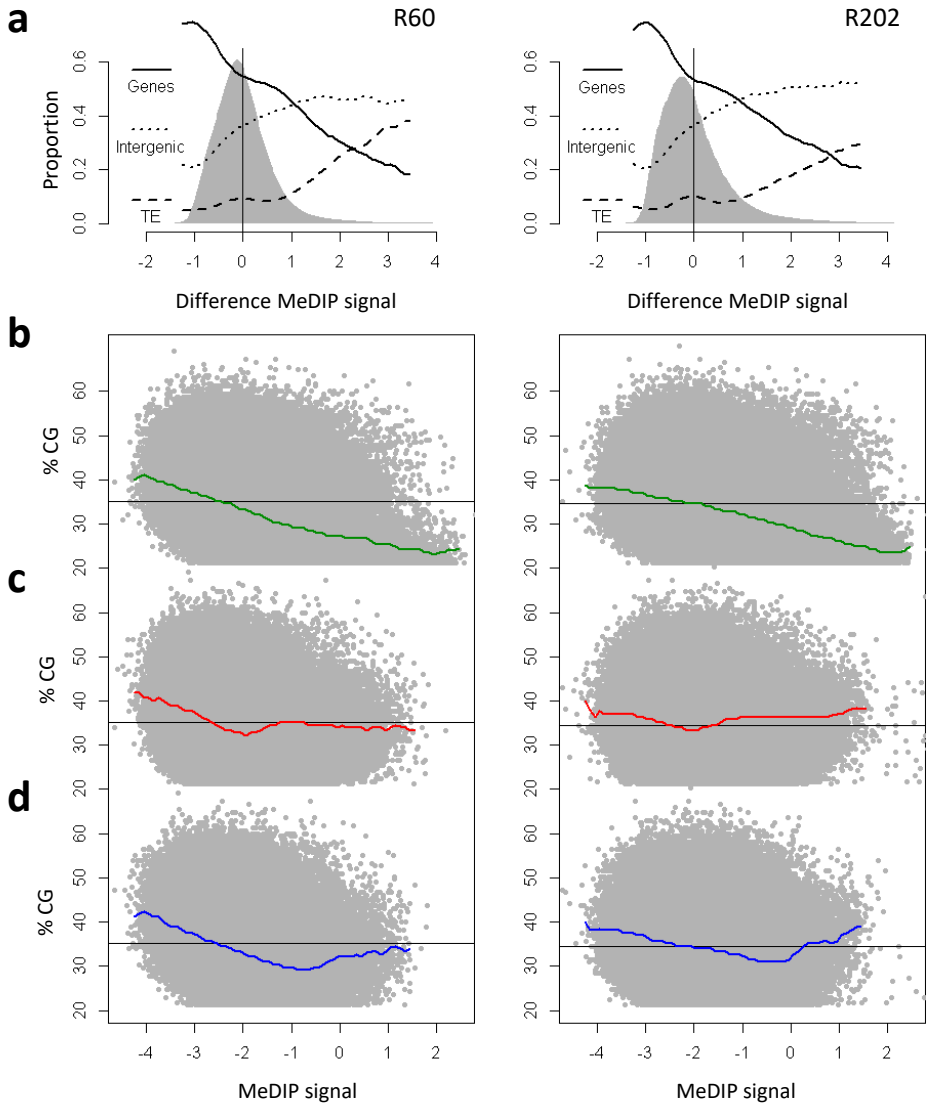

Figure 11. Annotation analysis and CG bias of unmethylated probes. (**a**) Shown is the proportion of genic, transposable element (TE) and intergenic probes along the entire MeDIP difference range (G/R–R/G). (**b**)–(**d**) Scatter plots of the CG percentage of each probe window (y-axis) and the MeDIP signal of each of the three MeDIP data sets (x-axis). The trend lines show a clear negative relationship between CG content and signal intensity. The color of each line corresponds to the three data sets (green: G/R data; red: R/G data; blue: DS data).

chip data where CG content is correlated with DNA methylation levels. Correcting for CG content, in this case, will reduce signal intensities arising from probes with true positive methylation measurements. A simple solution to bypass these issues is to work exclusively with R/G data where CG bias appears to be minimal (Fig. 11c, *see* **Note 2**).

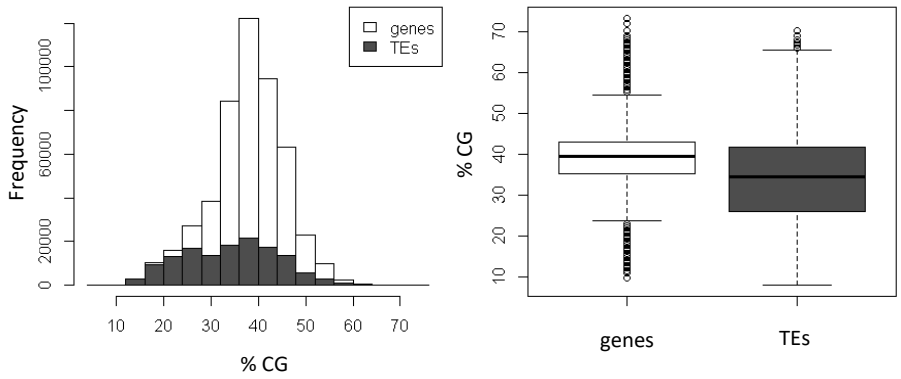

Figure 12. CG content of genes and transposons. Shown are the CG content distributions of genes and transposable elements (TEs).

2.4 Concluding remark

Although the mapping resolution of MeDIP-chip (~165 bps) is much lower than the single cytosine measurements that can be achieved with WGBS-seq, this array-based technology provides a level of resolution that should be sufficient for the detection of most functionally important differentially methylated regions. MeDIP-chip requires fewer bioinformatic resources and therefore scales more easily to large samples. Provided several experimental and data preparation steps are followed (*see* **Note 1–3**), MeDIP-chip presents a viable alternative to WGBS-seq in future population epigenetic studies.

2.5 Notes

1. *Recommendations for data preparation:* Prior to MeDIP-chip analysis, potentially cross-hybridizing probes should be removed. They typically show signal intensities similar to probes that correspond to genomic regions with high methylation levels (Fig. 3). Failure to remove cross-hybridizing probes can therefore result in the detection of a large number of false positives. However, removal of these probes will result in loss of measurement coverage; but this drawback is no different from sequencing-based approaches where short reads that do not map uniquely are usually excluded.
2. *Recommendations for dye-labeling:* Dye-related biases can pose serious concerns in dual-channel MeDIP-chip. Labeling the immunoprecipitated (IP) DNA with Cy3 (green) and the control DNA (input) with Cy5 (red) (i.e., G/R data) introduces strong signal distortions that significantly compromise measurement accuracies. These biases are particularly pronounced in genomic regions with low methylation and low CG content. This signal bias disappears when the opposite labeling strategy is employed (labeling IP with Cy5 and input with Cy3, i.e., R/G data). As a result of the G/R dye bias, dye-swap experiments in MeDIP-chip always perform worse than the R/G data alone, despite the fact that DS consists of twice as much data. Hence, despite its routine use in expression micro-array studies, we do not recommend the use of dye-swaps in dual-channel MeDIP-chip. This means that experimental costs can be reduced by a factor of two without loss of measurement information.
3. *Recommendations for data analysis:* The classification of probes as methylated or unmethylated requires a sound statistical approach. The best methods for MeDIP-chip are variants of HMMs [28]. The assumptions of HMMs are fundamentally consistent with the data properties arising from MeDIP-chip experiments. These assumptions are the following: (1) A probe signal is a noisy proxy for an underlying (unobserved) methylation state and (2) methylation states are spatially correlated along the genome owing to the array design and the propensity of DNA methylation to occur in clusters. Our application of a recent HMM designed for *Arabidopsis* MeDIP-chip [15] resulted in a genome-wide false-positive rate of about 0.02 and false-negative rate of about 0.19 (a combined rate of 0.21) for the R/G data. This relatively high false-negative rate implies that the application of this HMM misses

regions with low methylation levels. Less customized methods may yield even higher misclassifications. In population studies, this limitation will restrict the calling of DMRs to clear methylation differences between individuals (e.g., no methylation versus high methylation) and will likely fail to detect more subtle DMRs (e.g., no methylation versus low methylation). One way to improve this situation is to consider at least one additional technical R/G replicate.

Acknowledgments

This work was supported by grants from the Netherlands Organization for Scientific Research (NWO) (to F.J. and M.C.-T) and the Netherlands Bioinformatics Centre (NBIC) (to R.W.). Work in the Colot lab is supported in part by the European Union Network of Excellence EpiGeneSys.

References

1. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**:204–220.
2. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* **11**:191–203.
3. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126**:1189–1201.
4. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* **39**:61–69.
5. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**:523–536.
6. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulfite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**:215–219.

7. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334**:369–373.
8. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**:245–249.
9. Boyes J, Bird A (1992) Repression of genes by DNA methylation depends on CpG density and promoter strength: Evidence for involvement of a methyl-CpG binding protein. *EMBO J* **11**:327–333.
10. Lorincz MC, Schübeler D, Hutchinson SR, Dickerson DR, Groudine M (2002) DNA methylation density influences the stability of an epigenetic imprint and Dnmt3a/b-independent de novo methylation. *Mol Cell Biol* **22**:7572–7580.
11. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, Bouchez D, Dillmann C, Guerche P, Hospital F, Colot V (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5**:e1000530.
12. Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E, Labadie K, Wincker P, Jacobsen SE, Jansen RC, Colot V, Johannes F (2012) Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci USA* **109**:16240–16245.
13. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**:853–862.
14. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavaré S, Beck S (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26**:779–785.
15. Cortijo S, Wardenaar R, Colomé-Tatché M, Johannes F, Colot V (2014) Genome-wide analysis of DNA methylation in *Arabidopsis* using MeDIP-chip. *Methods Mol Biol* **1112**:125–149.
16. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA* **89**:1827–1831.

17. Clark SJ, Harrison J, Paul CL, Frommer M (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res* **22**:2990–2997.
18. Lister R, Ecker JR (2009) Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Res* **19**:959–966.
19. Krueger F, Kreck B, Franke A, Andrews SR (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat Methods* **9**:145–151.
20. Chen PY, Cokus SJ, Pellegrini M (2010) BS Seeker: Precise mapping for bisulfite sequencing. *BMC Bioinformatics* **11**:203.
21. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**:185–193.
22. R Development Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0, URL http://www.R-project.org.
23. Dobbin KK, Kawasaki ES, Petersen DW, Simon RM (2005) Characterizing dye bias in microarray experiments. *Bioinformatics* **21**:2430–2437.
24. Dombkowski AA, Thibodeau BJ, Starcevic SL, Novak RF (2004) Gene-specific dye bias in microarray reference designs. *FEBS Lett* **560**:120–124.
25. Martin-Magniette ML, Mary-Huard T, Bérard C, Robin S (2008) ChIPmix: Mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics* **24**:i181–i186.
26. Andrews SR (2007) ChIPMonk: Software for viewing and analysing ChIP-on-chip data. *BMC Syst Biol **1***(Suppl 1):P80.
27. Johannes F, Wardenaar R, Colomé-Tatché M, Mousson F, de Graaf P, Mokry M, Guryev V, Timmers HT, Cuppen E, Jansen RC (2010) Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* **26**:1000–1006.
28. Seifert M, Cortijo S, Colomé-Tatché M, Johannes F, Roudier F, Colot V (2012) MeDIP-HMM: Genome-wide identification of distinct DNA methylation states from high-density tiling arrays. *Bioinformatics* **28**:2930–2939.
29. Li W, Meyer CA, Liu XS (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **21**(Suppl 1):i274–i282.
30. Royce TE, Rozowsky JS, Gerstein MB (2007) Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics* **23**:988–997.
31. Gilbert D, Rechtsteiner A (2009) Comments on sequence normalization of tiling array expression. *Bioinformatics* **25**:2171–217

Chapter 3

Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation

---

Published as:

Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E, Labadie K, Wincker P, Jacobsen SE, Jansen RC, Colot V, Johannes F (2012) Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci USA* **109**:16240–16245.

Chapter 3

Abstract

The rate of meiotic crossing over (CO) varies considerably along chromosomes, leading to marked distortions between physical and genetic distances. The causes underlying this variation are being unraveled, and DNA sequence and chromatin states have emerged as key factors. However, the extent to which the suppression of COs within the repeat-rich pericentromeric regions of plant and mammalian chromosomes results from their high level of DNA polymorphisms and from their heterochromatic state, notably their dense DNA methylation, remains unknown. Here, we test the combined effect of removing sequence polymorphisms and repeat-associated DNA methylation on the meiotic recombination landscape of an *Arabidopsis* mapping population. To do so, we use genome-wide DNA methylation data from a large panel of isogenic epigenetic recombinant inbred lines (epiRILs) to derive a recombination map based on 126 meiotically stable, differentially methylated regions covering 81.9 % of the genome. We demonstrate that the suppression of COs within pericentromeric regions of chromosomes persists in this experimental setting. Moreover, suppression is reinforced within 3-Mb regions flanking pericentromeric boundaries, and this effect appears to be compensated by increased recombination activity in chromosome arms. A direct comparison with 17 classical *Arabidopsis* crosses shows that these recombination changes place the epiRILs at the boundary of the range of natural variation but are not severe enough to transgress that boundary significantly. This level of robustness is remarkable, considering that this population represents an extreme with key recombination barriers having been forced to a minimum.

3.1 Introduction

Meiotic recombination is a fundamental process in genetics whereby maternally and paternally inherited homologous chromosomes exchange material, either nonreciprocally by gene conversion or reciprocally by crossing over (CO). COs are not distributed uniformly along the genome but occur more often in chromosome arms and are strongly suppressed in pericentromeric regions [1–3], partly as a result of sequence and chromatin determinants [1, 4–8]. It is commonly believed that in plants and mammals high levels of DNA sequence polymorphisms as well as heterochromatic features associated with repeats, notably dense DNA methylation and transcriptional silencing, play a central role in this suppression [1, 4].

Suppression of COs by dense DNA methylation has been demonstrated experimentally in the fungus *Ascobolus* [7]. Specifically, COs were reduced when the

recombination interval was methylated on one homolog and were abolished almost completely when methylated on both homologs. In *Arabidopsis*, two recent mapping studies analyzed F2 progeny derived from crosses between Columbia *ddm1* and *met1* [Col(*ddm1*), Col(*met1*)] DNA methylation mutants and wild-type Landsberg [Ler(WT)] accessions and showed that loss of DNA methylation could not alleviate the suppression of COs in pericentromeric regions of chromosomes [9, 10]. However, as pointed out by the authors, this experimental design could not rule out an inhibitory effect of sequence divergence between Col and Ler on COs.

An ideal design would use crosses between isogenic individuals, with one of the crossing partners having decreased DNA methylation levels throughout the genome [9]. Melamed-Bessudo and Levy [9] implemented such an approach by crossing Col(*ddm1*) mutant to Col(WT). Using two fluorescent markers spanning a 16-centimorgan (cM) interval on the arm of chromosome 3, they detected increased CO rates in F2 plants derived from these parents relative to plants derived from a Col(WT) × Col(WT) control cross and concluded that COs in euchromatic regions can be up-regulated by loss of DNA methylation. A similar approach at a genome-wide scale and with high mapping resolution, particularly in pericentromeric regions, has not been attempted because of a lack of appropriate molecular and genetic tools. Hence, the combined effect of DNA methylation and sequence variation on COs has not been tested comprehensively in *Arabidopsis* or in any other higher eukaryote.

We previously reported the construction of a large population of epigenetic recombinant inbred lines (epiRILs) in *Arabidopsis* [11, 12], which provides a powerful experimental system to conduct such a test. These epiRILs were obtained by first crossing a fourth-generation plant homozygous for the recessive *ddm1-2* mutation with a near-isogenic WT individual. The *ddm1-2* mutation mostly affects transposable elements (TEs) and other repeats, which lose DNA methylation and become transcriptionally reactivated in a transmissible manner in many instances [11–14]. However, transposition events are relatively rare [15].

Thus, F1 individuals can be considered homozygous throughout the genome, except at the *DDM1* locus and at the few loci affected by TE mobilization, but have chromosome pairs that differ markedly in their DNA methylation levels and transcriptional activity over TEs and other repeats [11, 16]. A single F1 *DDM1/ddm1* individual was backcrossed to the WT parental line, and after the progeny homozygous for the WT *DDM1* allele were selected, the epiRILs were propagated through seven rounds of selfing. In this design, more than 85% of all informative recombination events occur in the first two inbreeding generations (F1 and backcross), with fewer informative events being contributed by each subsequent generation [17].

Previous targeted analysis indicated that many of the parental differences in DNA methylation and transcriptional activity of repeats are inherited stably in the epiRILs [11, 12]. Regions with segregating DNA methylation states therefore can serve as physical markers to detect the frequency and distribution of recombination events along chromosomes even though the two homologs have nearly identical DNA sequences.

In this study we report the construction of a recombination map using genome-wide DNA methylation data from 123 epiRILs. This map was derived from 126 meiotically stable differentially methylated regions (DMRs) covering 81.9 % of the total genome. Estimates of the genetic length for each chromosome revealed that global recombination rates are comparable with those of classical *Arabidopsis* crosses. On a local scale, we demonstrate that suppressed recombination activity within repeat-rich, pericentromeric regions of chromosomes is maintained robustly even after the removal of sequence polymorphisms and repeat-associated DNA methylation. Furthermore, we were able to identify 3-Mb regions flanking pericentromeric boundaries that appear to be subject to additional suppression and show that this effect is accompanied by increased recombination activity in chromosome arms. A direct comparison with 17 classical *Arabidopsis* crosses reveals that these recombination changes place the epiRILs at the boundary of the range of natural variation but appear not to be severe enough to transgress that boundary significantly.

3.2 Results

3.2.1 Construction of a recombination map using transgenerationally stable DMRs

To demonstrate that transgenerationally stable DMRs can be used for the construction of a recombination map in an isogenic population, we carried out methylated DNA immunoprecipitation followed by hybridization to a whole-genome DNA tiling array (MeDIP-chip) on 123 epiRILs and on the two parental lines (256 array experiments including replicates). The 123 epiRILs originally were chosen using a selective (epi)genotyping strategy for two uncorrelated complex traits, flowering time and root length. We used a three-state Hidden Markov Model (HMM) to classify tiling array signals into three underlying DNA methylation states [18]: unmethylated (U), intermediate methylation (I), or methylated (M). Benchmarking of these HMM calls against whole-genome bisulfite sequencing data (~30×) for six epiRILs confirmed that both the MeDIP protocol and the analysis method performed well (SI

Appendix, Fig. S1 and Table S1). Comparison of the two parental DNA methylomes revealed 2,611 DMRs representing clear instances of methylation loss in *ddm1* (transitions from M to U). These DMRs (median length: 1,211 bp; range: 318–24,624 bp) were distributed throughout the genome but, as expected, were more abundant in pericentromeric regions (Fig. 1A and SI Appendix, Table S2) [19].

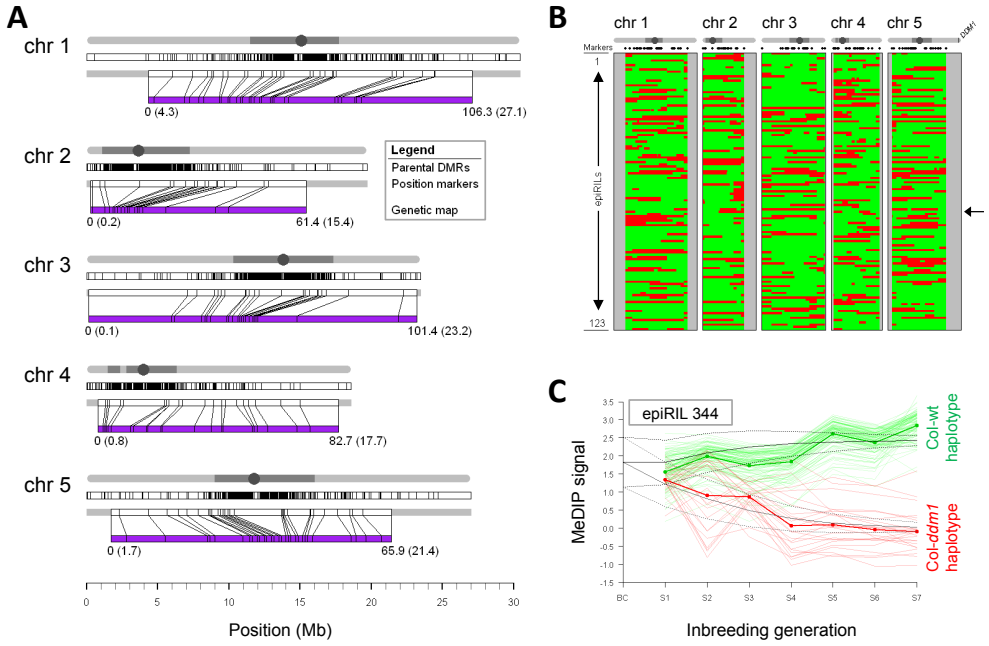

Figure 1. Recombination map construction. (**A**) Genome-wide distribution of the 2,611 parental DMRs (Top) and the 126 DMRs (i.e., markers; *Middle*) retained for construction of the recombination map (purple, *Bottom*) for each of the five *Arabidopsis* chromosomes. The mapping between physical and genetic positions of markers is shown. (**B**) Inference of inherited WT (green) and *ddm1* (red) haplotypes along the genome (x-axis) as inferred from the recombination map for each of the 123 epiRILs (y-axis) (SI Appendix, Table S5). Chromosome extremities not covered by the genetic map are indicated in gray. The genome of epiRIL 344 is indicated by an arrow. A schematic representation of each chromosome is plotted above the map with the physical location of the *DDM1* gene shown at the end of chromosome 5. (**C**) Transgenerational methylation data for epiRIL 344. Shown are the average methylation signals for the 126 markers, with regions that are predicted to become fixed for the *ddm1* haplotypes (thin red lines) and the WT haplotypes (thin green lines) after seven selfing generations. The average signals (red and green thick solid lines) are in agreement with Mendelian inbreeding theory (black solid lines).

We examined the DNA methylation state at all parental DMRs in each of the 123 epiRILs and inferred their parent of origin (i.e., epigenotypes). Segregation was not compatible with stable inheritance of *ddm1*-induced DNA hypomethylation for 1,744 (66.8 %) of the parental DMRs, and in most of these cases our data pointed to fully or partially penetrant reversion to WT DNA methylation. In contrast, 867 (33.2%) of the parental DMRs segregated in the expected 3:1 Mendelian ratio (SI Appendix, Fig. S2 and Table S3). Stable DMRs were associated with a comparatively lower abundance of siRNAs in the WT and *ddm1* parental lines (SI Appendix, Fig. S3). These findings are in agreement with previous analyses [11, 12] and indicated that the 867 stable DMRs are not efficient targets of siRNA-mediated DNA remethylation, even after eight rounds of meiosis. These stable DMRs therefore could serve as physical markers in an extension of the Lander–Green algorithm [20] to derive a genetic map. After application of the algorithm and removal of mainly genetically redundant markers (i.e., markers located less than 0.0001 cM apart), 126 of the original 867 markers were retained (Fig. 1A and B and SI Appendix, Fig. S2 and Table S4). These 126 markers covered ~81.9 % of the total genome (74.7, 77.0, 98.4, 91.1, and 73.0 % of chromosomes 1, 2, 3, 4, and 5, respectively).

Many of the 126 markers contained TE sequences, consistent with the targeted effect of *ddm1* on these and other repeats (SI Appendix, Fig. S4). However, in a vast majority of cases, markers included only TE relics, which likely have lost their capacity to be mobilized (SI Appendix, Table S6). Indeed, both comparative genomic hybridization (SI Appendix, Fig. S5) and preliminary whole-genome resequencing suggested that none of the 126 DMRs contain sequences that were mobilized in the parental *ddm1* line or the epiRILs (SI Appendix, Table S6). Consistent with this finding, pair-wise recombination fractions between the 126 markers indicated a well-behaved and robust genetic map, reminiscent of those typically seen in classical crosses involving DNA sequence markers, with high correlation among linked loci and virtually no correlations among loci in different chromosomes (SI Appendix, Fig. S6). Moreover, all inferred *ddm1*-inherited non-recombinant pericentromeric haplotypes contained significantly less DNA methylation and were more actively transcribed than their WT counterparts (SI Appendix, Figs. S7 and S8).

To test further the transgenerational stability of the 126 markers as well as our inference of the parental epigenotypes at these marker locations, we performed genome-wide DNA methylation analysis for one selected line (epiRIL 344) for each of its seven selfing generations (7 × 2 replicates = 14 array experiments). Fixation occurred for the predicted parental epigenotype in each case, and the rate of approach toward fixation was consistent with Mendelian inbreeding theory for a backcross-derived RIL [19] (Fig. 1C). Taken together, these results rule out any

ambiguity in the actual location or DNA methylation state of the stable DMRs used for constructing the genetic map.

3.2.2 Total genetic length in the epiRILs does not diverge significantly from the natural range

One approach for evaluating the epiRILs recombination map is by comparison with a Col(WT) × Col(WT)-derived reference cross. In this set-up, changes in recombination patterns can be attributed directly to DNA methylation loss. However, tracking recombination events in such a reference is experimentally challenging. It requires a system akin to the fluorescent marker reporters used by Melamed-Bessudo and Levy [9], which does not easily scale genome-wide. An alternative approach is to evaluate the epiRILs in the context of natural variation. In terms of DNA sequence and DNA methylome divergence of its founder parental lines, the epiRILs can be viewed as representing an extreme situation with key barriers to recombination having been forced to a minimum. An important question therefore is how genome-wide recombination patterns in this population compare with those seen in crosses derived from different pairs of natural accessions.

We estimated the genetic length for each of the five epiRIL chromosomes using Haldane's map function. The lengths were 106.3, 61.4, 101.4, 82.7, and 65.9 cM for chromosomes 1–5, respectively, and correlated positively with physical chromosome length (SI Appendix, Fig. S9). The total length of the genetic map was 417.7 cM, yielding an average marker spacing of ~0.804 Mb (3.45 cM). These estimates are similar to those previously reported for genetic maps based on classical *Arabidopsis* crosses [21–24]. The use of other map functions that account for CO interference, such as the Kosambi or Carter and Falconer functions, yielded very similar results (SI Appendix, Fig. S10). To perform a more direct comparison between the epiRIL map and those of classical *Arabidopsis* crosses, we reanalyzed recombination data obtained for 17 F2 populations [24] that were derived from pairs of 18 distinct natural accessions. In total, these populations consisted of 7,045 plants (~410 plants per cross; range: 235–462 plants), which were genotyped at 235 markers on average (range: 215–257 markers) [24]. To facilitate a meaningful comparison, we constructed a consensus map using 83 markers that were shared across populations (SI Appendix, Fig. S11 and Table S7). Thorough testing showed that the reduction to 83 markers in the epiRIL and F2 maps led to no significant loss of information in capturing the linkage structure along chromosomes (SI Appendix, Figs. S12 and S13), and the 83 markers therefore were deemed appropriate for this comparative analysis.

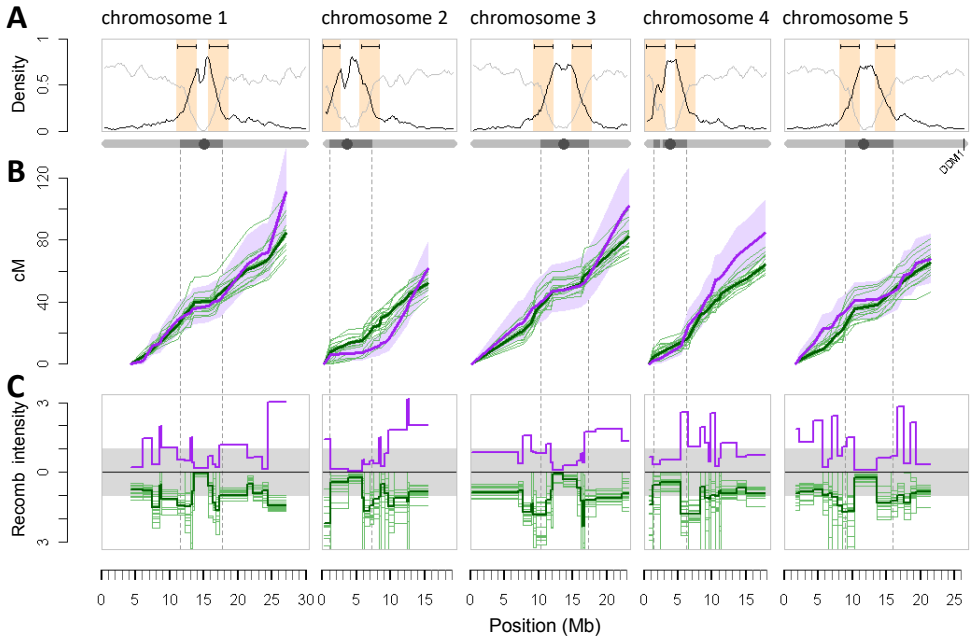

Figure 2. Comparison of global and local recombination patterns in the epiRILs and the 17 F2 populations [24]. (**A**) Chromosome-wide gene (light gray line) and transposon (dark line) density distribution. The 3-Mb windows bracketing the intersection points between transposon- and gene-dense regions are indicated in orange. (**B**) Cumulative cM lengths of the epiRILs (thick purple line) and each of the F2 populations (thin green lines) using the consensus map. Purple shading shows the ±95 % confidence interval (CI).The thick green line denotes the average F2 cumulative length (in cM). The dotted vertical lines define the pericentromeric regions of each chromosome. (**C**) The distribution of normalized recombination intensities (cM/Mb of a given marker interval divided by the cM/Mb chromosome average) shows suppression of recombination within pericentromeric regions and elevation at its boundaries. Color coding is as in **B**.

Estimates of the genetic length of each of the five chromosomes revealed substantial natural variation among the F2 populations (Figs. 2B and 3A). However, the genetic lengths of the epiRIL chromosomes did not diverge significantly from the natural range (Figs. 2B and 3A). The exception was chromosome 1, where we observed a significant increase relative to five of the F2 crosses. Overall, therefore, our data indicate that the global recombination rate in the epiRILs is not altered drastically. Nonetheless, we noted a clear, but non-significant, trend toward longer genetic

lengths for chromosomes 1–4 as compared with the F2 populations (Fig. 3A); this trend is at least partly consistent with DNA methylation and DNA sequence polymorphisms acting as barriers to the global recombination rate in *Arabidopsis*.

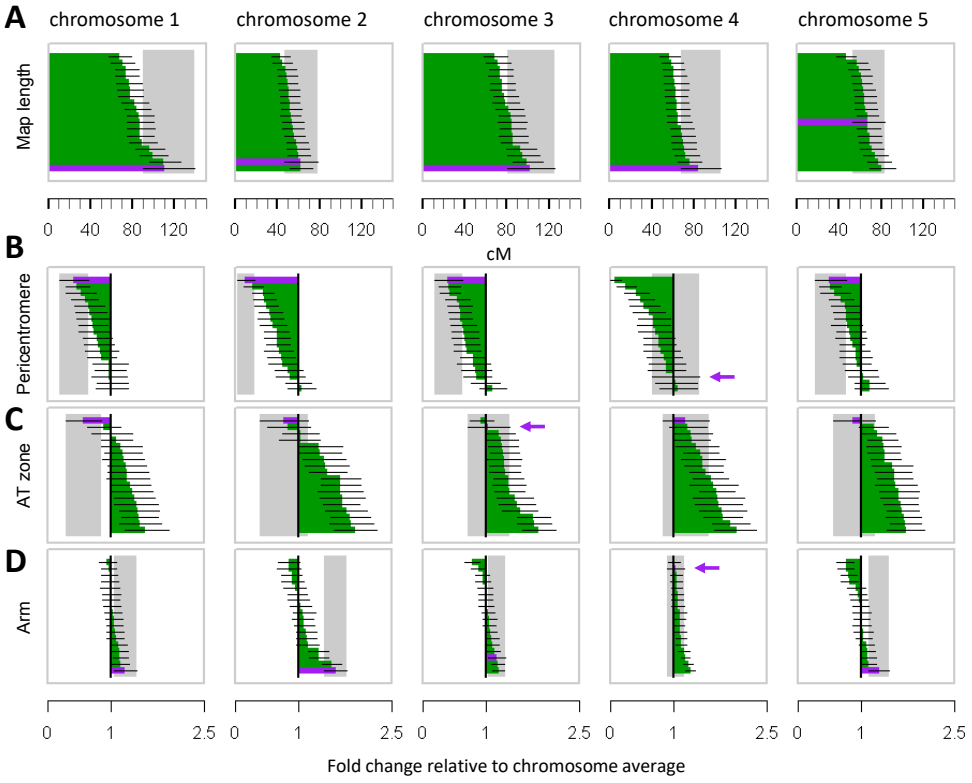

Figure 3. Estimated genetic lengths and fold-change recombination intensities. (**A**) Estimated genetic lengths (±95 % CI) of the epiRILs (purple) and each of the 17 F2 populations (green) [24]. (**B–D**) Fold-change in recombination intensity [(cM/Mb) region/(cM/Mb) chromosome average] ±95 % CI in pericentromeric regions (**B**), AT zones defined by a 3-Mb window bracketing the intersection point between transposon- and gene-dense regions at pericentromeric boundaries (**C**), and chromosome arms (**D**). Purple arrows indicate the location of the epiRILs when applicable. The values presented in each panel are ordered to highlight trends in the epiRILs recombination landscape. The identifiers of individual F2 crosses corresponding to this ordering can be found in SI Appendix, Table S11.

3.2.3 Suppression of pericentromeric recombination persists in the epiRILs and shows a trend toward additional reinforcement

To explore the relationship between the epiRILs map and those of the different F2 crosses at a subchromosomal scale, we examined in more detail the distribution of recombination intensities, expressed as cM/Mb, for each marker interval along the genome (Fig. 2C). All populations, including the epiRILs, had clearly suppressed recombination activity across pericentromeric regions relative to the chromosome averages (Figs. 2C and 3B). The exception to this trend was chromosome 4, for which the epiRILs showed a slight increase of recombination intensity (Fig. 3B). However, the presence of the heterochromatic knob on chromosome 4 in the Columbia accession, but not in other accessions, makes this result difficult to interpret [10].

Specifically, recombination intensities in pericentromeric regions of epiRIL chromosomes 1, 2, 3, and 5 were, respectively, 2.50, 6.88, 2.53, and 2.01 times lower than the chromosome average, which compares to 1.27 (range: 0.97–2.15), 1.51 (range: 0.95–3.68), 1.52 (range: 0.90–2.48), and 1.20 (range: 0.87–1.98) in the F2 populations (Fig. 3B and SI Appendix, Table S8). This persistent suppression effect in the epiRIL agrees with the results of Melamed-Bessudo and Levy [9] and Mirouze *et al.* [10], who examined mapping populations derived from a Col(*ddm1*) × Ler(WT) and a Col(*met1*) × Ler(WT) cross, respectively. Hence, loss of DNA methylation appears to be insufficient to release pericentromeric suppression of recombination, even in the absence of DNA sequence polymorphisms. On the contrary, we found a clear trend toward enhanced suppression in the epiRILs: Recombination intensities in this population were consistently at the bottom of the natural range compared with the F2 populations, even though chromosome-wide recombination rates were comparatively large. Enhanced suppressive effects also were reported by Melamed-Bessudo and Levy [9] and Mirouze *et al.* [10], thus highlighting an unexpected and complex relationship between DNA methylation and the suppression of recombination in pericentromeric regions of *Arabidopsis* chromosomes.

3.2.4 Reinforced suppression of recombination extends to pericentromeric boundaries in the epiRILs and appears to be compensated by increased recombination in chromosome arms

In contrast to core pericentromeric regions, recombination intensities in the F2 populations increase rapidly at pericentromeric boundaries with chromosome arms (Figs. 2C and 3C). An important property of these regions is that they correspond to major transitions in genome content from TE-rich to gene-rich sequences (Fig. 2A)

and also have been described recently as hotspots of historical recombination activity at the species level (SI Appendix, Fig. S14) [25]. We found that nearly 40 % of all detected recombination breakpoints in the F2 populations mapped within a 3-Mb window bracketing the intersection point in these transition zones (henceforth referred to as "annotation transition zones"; AT zones), yielding local recombination intensities that were consistently above the chromosome averages (Fig. 3C and SI Appendix, Fig. S15 and Table S8).

This finding differs strongly from the situation seen in the epiRILs: AT zones accounted for only 25.31% of all detected recombinants in this population, and recombination intensities were close to the chromosome average (in chromosomes 3 and 4) or even below it (in chromosomes 1, 2, and 5) (Fig. 3C and SI Appendix, Table S8). These results suggest that the enhanced suppression of recombination seen in the epiRIL pericentromeric regions (*see* above) is driven at least in part by the more localized reduction of recombination within AT zones, which cover (on average) only 63.4% of the pericentromeric regions on either side of the centromeres. The two previous studies using mapping populations derived from crosses between Col(*ddm1*) and Ler(WT) [9] and between Col(*met1*) and Ler(WT) [10] were not able to delineate these local effects, most likely because of the sparsity of their genetic markers (two to three markers per pericentromeric region). Marker density in the epiRIL map, in contrast, was relatively high within AT zones and even permitted fine mapping of shared and nonshared recombination breakpoints to a resolution as low as 4 kb (SI Appendix, Figs. S16 and S17 and Tables S9 and S10).

Furthermore, our data indicate that suppression of recombination within AT zones in the epiRILs is accompanied by increased recombination in chromosome arms (Fig. 3D and SI Appendix, Fig. S18). This apparent compensatory effect reconciles the enhanced local suppression seen in the epiRILs with the earlier observation that chromosome-wide recombination rates are relatively large compared with the F2 populations. This effect was most pronounced on epiRIL chromosomes 1, 2, and 5 (the chromosomes with the strongest suppression in the AT zone), with recombination intensities being 1.23, 1.6, and 1.3 times above the chromosomes' average (SI Appendix, Table S8). We failed to identify a similar trend in the F2 populations (SI Appendix, Fig. S18 and Table S8), suggesting that this effect is a specific feature of the epiRIL recombination landscape.

3.3 Discussion

In this study we demonstrate that stable DNA methylation differences can be used as physical markers to derive genome-wide recombination patterns in a near

isogenic population of epiRILs. We find that recombination suppression is maintained robustly in pericentromeric regions of the epiRILs, despite the extensive loss of sequence variation and of DNA methylation and transcriptional silencing over repeats. This observation indicates that these factors do not play a major role in the suppression of pericentromeric COs. This finding is contrary to common belief and is particularly intriguing given the interplay between recombination and transcription observed in yeast and the mouse [26, 27]. Whether mechanisms exist in *Arabidopsis* that actively sequester the recombination machinery away from gene-promoter regions or other genomic elements, as in the mouse [27], remains to be determined.

Nonetheless, our results indicate that loss of DNA methylation over repeat sequences can lead to a local reinforcement of recombination suppression in pericentromeric regions and to increased recombination activity along chromosome arms. Similar results were reported by Melamed-Bessudo and Levy [9] and Mirouze *et al.* [10] using genetically divergent populations. Therefore we conclude that the absence of sequence polymorphisms is insufficient to counteract the enhanced suppressive effects induced by the loss of DNA methylation in pericentromeric regions. On the other hand, the lack of sequence polymorphisms still may be partly responsible for the increased recombination rates observed in chromosome extremities [9].

Melamed-Bessudo and Levy [9] demonstrated that *ddm1*-induced demethylation of only one homolog produces the same recombination changes seen when both homologs are demethylated. Our results and conclusions therefore should be generalizable to the two-homolog situation. However, it has been shown in *Ascobolus* that DNA methylation of a known recombination hotspot inhibits COs more severely when both homologs are methylated [7]. Similar localized dosage effects may therefore also be present in *Arabidopsis*.

Our study and those of Melamed-Bessudo and Levy [9] and Mirouze *et al.* [10] have used well-characterized *ddm1* and *met1* DNA methylation mutants as a tool to perturb genome-wide methylation levels experimentally. Both *ddm1* and *met1* experience a nearly 70 % reduction in DNA methylation levels genome-wide. This drastic loss probably sets an upper limit to the amount of demethylation that can be incurred in nature. Indeed, it is difficult to conceive of mechanisms that would elicit similar or more severe changes under natural settings, unless they involve spontaneous mutations in genes important for DNA methylation control, such as *ddm1* or *met1*. Interestingly, a recent analysis of *Arabidopsis* mutation accumulation lines showed that drastic alterations in the methylome of one outlier line were likely caused by a spontaneous mutation in a methyltransferase gene [28, 29], which must

have arisen during just 30 generations of selfing. This observation suggests that similar events are certainly plausible under natural conditions.

An assessment of whether strong methylation loss can elicit recombination changes at magnitudes that are sufficient to drive genome evolution in this species has been lacking. Our study is an initial step in providing such an assessment. Our analysis of the 17 F2 populations derived from 18 natural accessions [24] allowed us to quantify the magnitude of the recombination changes observed in the epiRILs in the context of natural variation. Although we find that the epiRILs nearly always are situated at the boundary of the natural range, there is no strong evidence that local and global recombination patterns in this population markedly transgress the natural range. Indeed, in many cases, several of the F2 populations displayed even more extreme divergence from the F2 population average than did the epiRILs. These findings lead us to conclude that severe losses of DNA methylation along *Arabidopsis* chromosomes have no drastic implications for recombination-mediated genome evolution. This high level of robustness raises questions concerning the precise mechanisms that have shaped the recombination landscape in this species in the first place.

Of course, severe depletion of DNA methylation can drive other important events, such as large-scale structural rearrangements and polyploidization, which may impact the course of genome evolution. In addition, natural epigenetic variation, such as that associated with differential DNA methylation, can act on complex traits that are under natural selection [30], thereby changing linkage disequilibrium relations within and across chromosomes. However, understanding and documenting the impact of epigenetic variants on complex traits is challenging, mainly because of the technical difficulties in ruling out the confounding effect of DNA sequence polymorphisms [31]. Because of this limitation, it has been argued that the epiRILs constitute an ideal system for the study of epigenetic inheritance in *Arabidopsis* [17, 32–34]. We and others have shown recently that many adaptive phenotypes, such as plant height, flowering time, and growth rate, are highly heritable in this population [12, 35, 36]. Segregating phenotypic effects also have been observed in another epiRIL population which was obtained from a cross between Col(*met1*) and Col(WT) [37].

A logical next step in the analysis of these populations is to map and characterize the epigenetic basis of these complex traits. The linkage map reported here (Fig. 1B) can be used in conjunction with classical quantitative trait-locus mapping methods to achieve this characterization in the *ddm1*-derived epiRILs. Ultimately, such efforts should contribute significantly to our understanding of epigenetics in adaptive evolution.

3.4 Materials and methods

3.4.1 Methylome analysis

MeDIP was carried out as previously described [18] followed by hybridization to a custom NimbleGen tiling array covering the *Arabidopsis* genome at 165 nt resolution [38]. Including dye-swaps, we performed a total of 256 array experiments (SI Appendix, section 1). For each array, probe signals were classified into three underlying methylation states, methylated (M), intermediate (I), or unmethylated (U), using the HMM presented previously (SI Appendix, section 2) [18]. These inferred methylation states were cross-validated against whole-genome bisulfite sequencing data of six epiRILs (SI Appendix, section 3, Fig. S1, and Table S1).

3.4.2 Definition of parental DMRs

We conducted a probe-level comparison of the HMM calls between the *ddm1* and WT parents (SI Appendix, section 4). Probe-level methylation calls were denoted as polymorphic when the parents differed (e.g., I in *ddm1* and M in WT) and as nonpolymorphic when they were identical (e.g., U in *ddm1* and U in WT). Neighboring probes reporting the same polymorphic state were collapsed into single regions. Hence, parental DMRs were defined as regions of at least three consecutive probes that reported the same extreme polymorphic state (i.e., transitions from M in WT to U in *ddm1* or vice versa). We found 2,611 DMRs, all of which were U in the *ddm1*. Detailed summary statistics are given in SI Appendix, Table S2.

3.4.3 Calling of parental origin of DMRs in the epiRILs

For any given epiRIL the parental origin of each DMR (i.e., epigenotype) was determined using an HMM-based inference method (SI Appendix, section 4).

3.4.4 Mendelian segregation criterion

Under the assumption that DMRs were stable for eight generations of breeding, both WT- and *ddm1*-like parental states should appear according to Mendelian segregation ratios in the epiRILs. The sampling variation around these ratios was calculated from a binomial distribution taking into account the sample size (n = 123), the cross design, and the 8 % F2 contamination previously reported [12]. DMRs in the epiRILs showing a percentage of WT-like states between 62.7 % and 83.3 % (the

expected value being 73 %) were taken as putative transgenerationally stable markers. In total 867 parental DMRs fulfilled this criterion and were used subsequently as a starting point for map construction (SI Appendix, section 5, Fig. S2, and Table S3).

3.4.5 Extension of Lander–Green algorithm

Derivation of a genetic map using DMRs was carried out through an extension of the Lander-Green algorithm [20], which was designed to accommodate marker and individual specific error rates. Our implementation of this algorithm is detailed in SI Appendix, section 6.

3.4.6 Transcriptome analysis of epiRILs and *ddm1* seedlings

Whole-genome expression profiling was performed using a custom NimbleGen tiling array [37]. For experimental details, *see* SI Appendix, section 7.

3.4.7 Transgenerational analysis of DMRs

MeDIP-chip was carried out for epiRIL 344 for seven generations of selfing after the backcross, following the protocol described above. At each generation, DNA from five siblings was pooled. The expected signal behavior was derived using a Markov Chain strategy, considering the Mendelian inheritance of the marker probes (SI Appendix, section 8).

3.4.8 Construction of the consensus map

To facilitate a meaningful comparison of the epiRILs map with those of the 17 different F2 populations, we constructed a consensus map (SI Appendix, section 9 and Fig. S11) by using the epiRILs map as a reference and selecting from each of the F2 maps the SNPs closest to the reference, allowing a maximum distance of ±1.39 Mb. The average distance from reference was ±0.17 Mb, which led to little loss of information in capturing the recombination structure along the genome (SI Appendix, Figs. S12 and S13). Markers deemed too distant were not included in the consensus map. This process resulted in 83 markers (SI Appendix, Table S7).

3.4.9 Recombination intensities at major annotation transitions

Fig. 2A and C shows that the recombination intensity increases rapidly at the pericentromeric boundaries, which also coincide with major transitions in genome content from genes to transposons. To find the area where the recombination intensity is maximal, we implemented a sliding window approach (SI Appendix, section 10 and Fig. S15).

3.5 Note added in proof

During the reviewing process, Yelina *et al.* (Yelina NE, Choi K, Chelysheva L, Macaulay M, de Snoo B, *et al.* (2012) Epigenetic remodeling of meiotic crossover frequency in *Arabidopsis thaliana* DNA methyltransferase mutants. *PLoS Genet* **8**: e1002844) reported elevated centromere-proximal COs, coincident with pericentromeric decreases and distal increases in *met1* mutants. However, total numbers of CO events were found to be similar between wild-type and *met1*. These results support the trends observed in the epiRIL population.

Acknowledgments

We thank Tony Heitkam for help with the TE analysis. This work was funded by grants from the Ministère de la Recherche et de l'Enseignement Supérieur (to S.C., B.L., and M.E.); Agence Nationale de La Recherche TAG and MEIOMETH projects (to V.C.) and EPIMOBILE project (to V.C. and P.W.); European Union Seventh Framework Programme Network of Excellence EpiGeneSys (Award 257082; to VC); Netherlands Organization for Scientific Research (NWO) (to F.J., M.L., and M.C.-T.); Consortium for Improving Plant Yield (CIPY) (to M.C.-T.); Netherlands Bioinformatics Centre (NBIC) (to R.W.); and EURATRANS (to R.C.J.). Work in the S.E.J. laboratory is supported by National Institutes of Health Grant GM60398. S.F. is a Special Fellow of the Leukemia & Lymphoma Society. S.E.J. is an investigator of the Howard Hughes Medical Institute.

Supplementary material

Supplementary text, figures and tables can be found at
http://www.pnas.org/content/109/40/16240.long?tab=ds.

References

1. Lichten M, de Massy B (2011) The impressionistic landscape of meiotic recombination. *Cell* **147**:267–270.
2. Mézard C, Vignard J, Drouaud J, Mercier R (2007) The road to crossovers: Plants have their say. *Trends Genet* **23**:91–99.
3. Muyt AD, Mercier R, Mézard C, Grelon M (2009) Meiotic recombination and crossovers in plants. *Genome Dyn* **5**:14–25.
4. Edlinger B, Schlögelhofer P (2011) Have a break: Determinants of meiotic DNA double strand break (DSB) formation and processing in plants. *J Exp Bot* **62**:1545–1563.
5. Chen W, Jinks-Robertson S (1999) The role of the mismatch repair machinery in regulating mitotic and meiotic recombination between diverged sequences in yeast. *Genetics* **151**:1299–1313.
6. Emmanuel E, Yehuda E, Melamed-Bessudo C, Avivi-Ragolsky N, Levy AA (2006) The role of *AtMSH2* in homologous recombination in *Arabidopsis thaliana*. *EMBO Rep* **7**:100–105.
7. Maloisel L, Rossignol JL (1998) Suppression of crossing-over by DNA methylation in *Ascobolus*. *Genes Dev* **12**:1381–1389.
8. Shi J, Wolf SE, Burke JM, Presting GG, Ross-Ibarra J, Dawe RK (2010) Widespread gene conversion in centromere cores. *PLoS Biol* **8**:e1000327.
9. Melamed-Bessudo C, Levy AA (2012) Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in *Arabidopsis*. *Proc Natl Acad Sci USA* **109**:E981–E988.
10. Mirouze M, Lieberman-Lazarovich M, Aversano R, Bucher E, Nicolet J, Reinders J, Paszkowski J (2012) Loss of DNA methylation affects the recombination landscape in *Arabidopsis*. *Proc Natl Acad Sci USA* **109**:5880–5885.
11. Teixeira FK, Heredia F, Sarazin A, Roudier F, Boccara M, Ciaudo C, Cruaud C, Poulain J, Berdasco M, Fraga MF, Voinnet O, Wincker P, Esteller M, Colot V (2009) A role for RNAi in the selective correction of DNA methylation defects. *Science* **323**:1600–1604.
12. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, Bouchez D, Dillmann C, Guerche P, Hospital F, Colot V (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5**:e1000530.
13. Vongs A, Kakutani T, Martienssen RA, Richards EJ (1993) *Arabidopsis thaliana* DNA methylation mutants. *Science* **260**:1926–1928.

14. Kakutani T, Munakata K, Richards EJ, Hirochika H (1999) Meiotically and mitotically stable inheritance of DNA hypomethylation induced by *ddm1* mutation of *Arabidopsis thaliana*. *Genetics* **151**:831–838.
15. Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T (2009) Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* **461**:423–426.
16. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**:471–476.
17. Johannes F, Colomé-Tatché M (2011) Quantitative epigenetics through epigenomic perturbation of isogenic lines. *Genetics* **188**:215–227.
18. Cortijo S, Wardenaar R, Colomé-Tatché M, Johannes F, Colot V (2014) Genome-wide analysis of DNA methylation in *Arabidopsis* using MeDIP-chip. *Methods Mol Biol* **1112**:125-149.
19. Bernatavichute YV, Zhang X, Cokus S, Pellegrini M, Jacobsen SE (2008) Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in *Arabidopsis thaliana*. *PLoS One* **3**:e3156.
20. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* **84**:2363–2367.
21. Giraut L, Falque M, Drouaud J, Pereira L, Martin OC, Mézard C (2011) Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. *PLoS Genet* **7**:e1002354.
22. Drouaud J, Mercier R, Chelysheva L, Bérard A, Falque M, Martin O, Zanni V, Brunel D, Mézard C (2007) Sex-specific crossover distributions and variations in interference level along *Arabidopsis thaliana* chromosome 4. *PLoS Genet* **3**:e106.
23. Drouaud J, Camilleri C, Bourguignon PY, Canaguier A, Bérard A, Vezon D, Giancola S, Brunel D, Colot V, Prum B, Quesneville H, Mézard C (2006) Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". *Genome Res* **16**:106–114.
24. Salomé PA, Bomblies K, Fitz J, Laitinen RA, Warthmann N, Yant L, Weigel D (2012) The recombination landscape in *Arabidopsis thaliana* F2 populations. *Heredity (Edinb)* **108**:447–455.
25. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, Muliyati NW, Platt A, Sperone FG, Vilhjálmsson BJ, Nordborg M, Borevitz JO, Bergelson J (2012) Genome-wide patterns of genetic variation in worldwide

Arabidopsis thaliana accessions from the RegMap panel. *Nat Genet* **44**:212–216.
26. Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG, Tischfield SE, Zhu X, Neale MJ, Jasin M, Socci ND, Hochwagen A, Keeney S (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* **144**:719–731.
27. Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV (2012) Genetic recombination is directed away from functional genomic elements in mice. *Nature* **485**:642–645.
28. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334**:369–373.
29. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**:245–249.
30. Richards EJ (2008) Population epigenetics. *Curr Opin Genet Dev* **18**:221–226.
31. Johannes F, Colot V, Jansen RC (2008) Epigenome dynamics: A quantitative genetics perspective. *Nat Rev Genet* **9**:883–890.
32. Richards EJ (2009) Quantitative epigenetics: DNA sequence variation need not apply. *Genes Dev* **23**:1601–1605.
33. Weigel D (2012) Natural variation in *Arabidopsis*: From molecular genetics to ecological genomics. *Plant Physiol* **158**:2–22.
34. Schmitz RJ, Ecker JR (2012) Epigenetic and epigenomic variation in *Arabidopsis thaliana*. *Trends Plant Sci* **17**:149–154.
35. Roux F, Colomé-Tatché M, Edelist C, Wardenaar R, Guerche P, Hospital F, Colot V, Jansen RC, Johannes F (2011) Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. *Genetics* **188**:1015–1017.
36. Latzel V, Zhang Y, Karlsson Moritz K, Fischer M, Bossdorf O (2012) Epigenetic variation in plant responses to defense hormones. *Ann Bot* **110**:1423-1428.
37. Reinders J, Wulff BB, Mirouze M, Marí-Ordóñez A, Dapp M, Rozhon W, Bucher E, Theiler G, Paszkowski J (2009) Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes Dev* **23**:939–950.
38. Roudier F, Ahmed I, Bérard C, Sarazin A, Mary-Huard T, Cortijo S, Bouyer D, Caillieux E, Duvernois-Berthet E, Al-Shikhley L, Giraut L, Després B, Drevensek S, Barneche F, Dèrozier S, Brunaud V, Aubourg S, Schnittger A, Bowler C, Martin-Magniette ML, Robin S, Caboche M, Colot V (2011)

Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. *EMBO J* **30**:1928–1938.

Chapter 4

Mapping the epigenetic basis of complex traits

Published as:

Cortijo S*, Wardenaar R*, Colomé-Tatché M*, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury JM, Wincker P, Roudier F, Jansen RC, Colot V, Johannes F (2014) Mapping the epigenetic basis of complex traits. *Science* **343**:1145-1148.

*Equal contribution

Abstract

Quantifying the impact of heritable epigenetic variation on complex traits is an emerging challenge in population genetics. Here, we analyze a population of isogenic *Arabidopsis* lines that segregate experimentally induced DNA methylation changes at hundreds of regions across the genome. We demonstrate that several of these differentially methylated regions (DMRs) act as bona fide epigenetic quantitative trait loci (QTL^{epi}), accounting for 60 to 90% of the heritability for two complex traits, flowering time and primary root length. These QTL^{epi} are reproducible and can be subjected to artificial selection. Many of the experimentally induced DMRs are also variable in natural populations of this species and may thus provide an epigenetic basis for Darwinian evolution independently of DNA sequence changes.

4.1 Introduction

Methylation of cytosines is an epigenetic mark involved in the silencing of transposable elements (TEs) and genes [1]. Despite its functional conservation across many species [2, 3], intraspecific surveys have revealed widespread variation in DNA methylation patterns within populations [4–6]. Estimates in the model plant *Arabidopsis thaliana* indicate that heritable changes in the methylation status of clusters of cytosines, which could be functionally more relevant than individual cytosines [7], arise spontaneously at rates similar to that of DNA sequence mutations [8, 9]. A key challenge in population genetics is to show that epigenetic variants exist independently of cis- or trans-acting DNA sequence changes, are stably transmitted over many sexual generations, and are associated with heritable phenotypic variation [10]. Addressing this challenge using natural populations continues to pose major technical difficulties.

To overcome these difficulties, we established in *Arabidopsis* a population of so-called epigenetic recombinant inbred lines (epiRILs) that have almost identical DNA sequences but segregate many differences in DNA methylation [11]. To derive this population, a plant homozygous for the recessive *ddm1-2* mutation was first crossed with a near-isogenic wild-type (WT) individual. The *ddm1-2* mutation leads to a loss of DNA methylation and silencing over transposable elements (TEs) mainly, with potential consequences on the expression of neighboring genes [12], but transposition events are relatively rare [13]. Importantly, some of the DNA methylation and expression changes induced by *ddm1-2* are inherited independently of the mutation [14, 15]. A single F1 *DDM1*/*ddm1-2* individual was therefore

backcrossed to the WT parental line, and after selection of F2 progeny homozygous for the *DDM1* allele, the epiRILs (N > 500) were selfed for six generations (Fig. S1).

Phenotypic analysis revealed significant broad-sense heritability in the epiRILs, with estimates ranging from about 0.05 to 0.4 [11, 16, 17]. Theoretical predictions indicate that these heritability values are consistent with a small number of parentally derived quantitative trait loci (QTL) [17, 18]. We hypothesized that these QTL are caused by stably inherited DNA methylation changes originating from the *ddm1-2* mutant founder parent. Indeed, a survey of the DNA methylomes of a selected set of 123 epiRILs identified hundreds of parental differentially methylated regions (DMRs) showing Mendelian segregation patterns [19]. Using an informative subset of 126 of these DMRs as physical markers, we were able to derive a genetic map covering 81.9% of the total genome [19].

Here, we used this map in conjunction with classical linkage analysis to search for epigenetic quantitative trait loci (QTL^{epi}) underlying complex traits in the epiRIL population.

4.2 Results

4.2.1 Interval mapping

We interval-mapped [20] two highly heritable (and weakly correlated) complex traits, flowering time (FT1) and primary root length (RL1) (Fig. 1, Fig. S2, and Table S1). The FT1 phenotype was obtained in a greenhouse experiment [11], whereas RL1 was measured in a climate-controlled growth chamber [21]. Linkage analysis detected highly significant QTL for FT1 on chromosome 1 (chr 1) [40.59 cM; logarithm of the odds ratio for linkage (LOD) = 8.72], chr 4 (30 cM; LOD = 4.43), and chr 5 (41.73 cM; LOD = 8.53) (Fig. 1A and Table S2). For these three QTL, the plants that inherited the WT epigenotype at the peak marker flowered significantly later than those with the *ddm1-2* inherited epigenotype (Fig. S3A). The combined additive effects of these QTL explained 86.78% of the broad-sense heritability for the trait and 51.14% of the total phenotypic variance (Fig. 1B and Table S3). All three QTL were confirmed using independent flowering time data (FT2) collected in a field experiment [17] (Fig. 1, A and B, and Table S3), which indicates that these QTL are robust across environmental settings. For RL1, we detected significant QTL on chr 1 (38 cM; LOD = 4.9), chr 2 (6.47 cM; LOD = 5.28), and chr 4 (50 cM; LOD = 2.65) (Fig. 1C and Table S2). The WT-inherited epigenotype at the peak QTL markers was associated with longer primary roots compared with the *ddm1-2* inherited epigenotype (Fig. S3B). The combined additive effect of these QTL explained 59.36%

of the estimated broad-sense heritability and 32.69% of the total phenotypic variance (Fig. 1D and Table S4). Again, all three QTL were confirmed with data from a replicate phenotyping experiment (RL2) (Fig. 1, C and D, and Table S4).

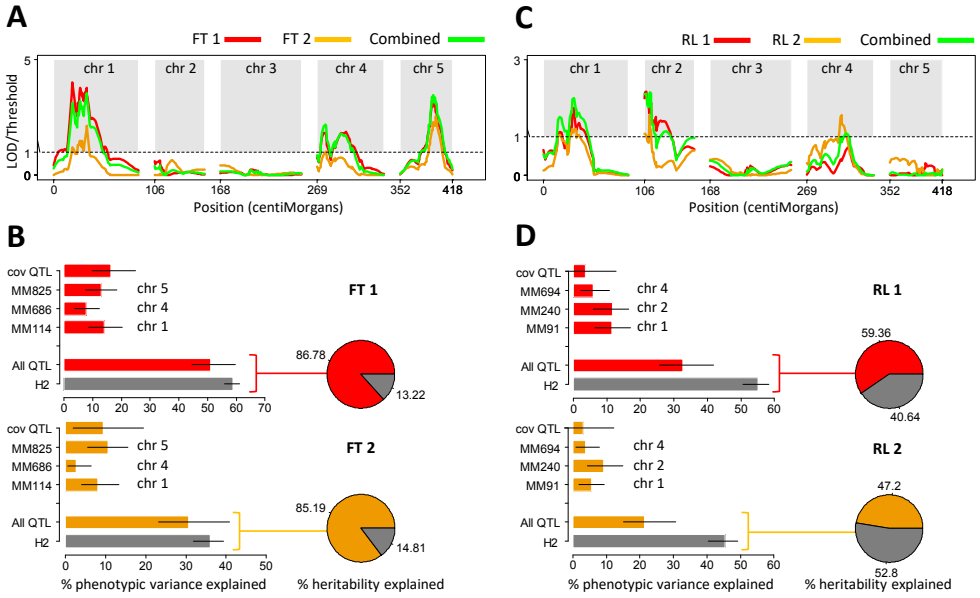

Figure 1. Interval mapping results. (**A**) QTL mapping profiles for two independent flowering time measurements (FT1 and FT2), as well as their average (combined). FT1 was measured in the greenhouse and FT2 in a field experiment. (**B**) Percentages of phenotypic variance and of broad-sense heritability (H^2) explained by the peak QTL markers (MM#). Error bars, ±1 standard error of the estimate. (**C**) QTL mapping profiles for two independent primary root length measurements (RL1 and RL2), as well as their average (combined). Both RL1 and RL2 were measured in a climate-controlled growth chamber. (**D**) Same as in (**B**), but for RL.

4.2.2 Ruling out *ddm1-2*-derived TE insertions

Our linkage mapping results indicate that the broad-sense heritability in the epiRILs is mainly due to causal variants originating from the parental generation and not from later generations of inbreeding. To examine the possibility that these parentally derived causal variants are transposable element (TE) insertions that occurred in the *ddm1-2* parental line rather than DMRs, we resequenced a representative sample of

52 of the 123 epiRILs (Tables S5 and S6). Our analysis revealed, in addition to several nonshared TE insertions, a total of four shared TE insertions in the RL and FT QTL

A

Trait	QTL			Inserted sequence			Number of epiRILs shared			Phenotypic effect (*p*-value)	
	Chr	Start	Stop	Name	Start	Stop	Seq	PCR	Total	TE	QTL marker
FT1	1	15908858	17261949	ATENSPM3	16836944	16837131	5 / 52	2 / 27	7 / 79	0.0076	5.20E-07
RL1	1	13617271	17523612	ATENSPM3	16836944	16837131	5 / 52	2 / 27	7 / 79	0.023	0.00012
RL1	1	13617271	17523612	ATENSPM3 / HELITRON	17407431	17410302	2 / 52	0 / 27	2 / 79	0.75	0.00012
RL1	4	8906583	11824662	ATCOPIA93	9649457	9651158	13 / 52	9 / 27	22 / 79	0.073	0.020
RL1	4	8906583	11824662	ATCOPIA78	10736583	10740627	18 / 52	8 / 27	26 / 79	0.99	0.020

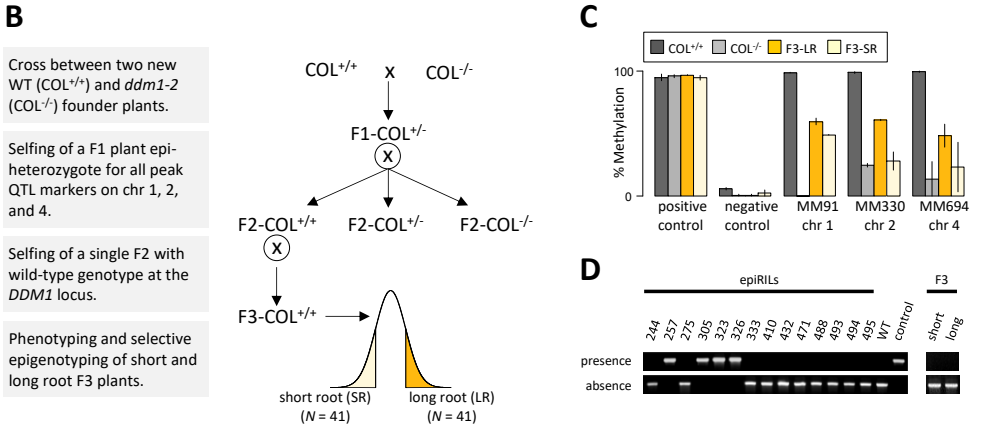

Figure 2. Ruling out *ddm1-2*-derived TE insertions as a cause for the epiRIL QTL. (**A**) Resequencing of 52 epiRILs and targeted PCR of an additional 27 epiRILs detected four shared insertions in the RL and FT QTL intervals (coordinates are according to TAIR10). Phenotypic analysis testing for the effect of TEs and peak QTL markers [*P* values from multiple regression models (Table S8)]. (**B**) Validation of the RL QTL by selective epigenotyping of 82 short- and long-root F3 progeny obtained using the crossing scheme shown. (**C**) Quantitative PCR analysis of McrBC-digested DNA of tail-selected samples was used to determine DNA methylation levels at the peak markers MM91, MM330, and MM694 on chr 1, 2, and 4, respectively (right panel). We used marker MM330 on chr 2 instead of the peak marker MM240. These markers are in tight linkage, but MM330 was easier to assay by PCR. Error bars, ±1 SEM. (**D**) Example of the presence or absence of the most common insertion (ATCOPIA 93) in a sample of epiRILs and the pools of short- and long-root F3 individuals. (Top) Results of PCR with primer pairs designed to amplify one end of the element and its flanking sequence. (Bottom) Results of PCR with primer pairs designed to amplify the WT sequence. Positive control: epiRIL 55.

confidence intervals (Fig. 2A, Fig. S4, and Table S6): two shared TE insertions in the chr 1 interval and two in the chr 4 interval. We were able to confirm the shared TE insertions using targeted polymerase chain reaction (PCR) assays in an additional 27 epiRILs (Fig. 2, A and D, and Tables S5 to S7).

However, further analysis revealed that shared TE insertions were not consistently inherited from the *ddm1-2* parent (Fig. S4 and Table S6). Thus, we found that the ATCOPIA78 insertion on chr 4 lies in an interval of WT origin and is present in about one-third of the epiRILs, which indicates that this insertion occurred in the F1 individual rather than in the *ddm1-2* parent. Similarly, the two ATENSPM3 insertions on chr 1 lie in intervals of *ddm1-2* origin but are absent in some epiRILs with this epigenotype, suggesting that they either occurred in the parental *ddm1-2* line, with excisions in some epiRILs, or else arose in the F1 individual. Finally, the ATCOPIA93 insertion on chr 4 lies in an interval that is of *ddm1-2* origin in some epiRILs and of WT origin in others, reflecting a highly dynamic inheritance pattern. Attempts to associate these shared TE insertions with phenotypes revealed a significant association for ATENSPM3 (chr 1) with root length and flowering time (FT, $P = 0.0076$; RL, $P = 0.023$) (Fig. 2A and Table S8A) and a borderline significant effect for the ATCOPIA93 (chr 4) insertion on primary root length ($P = 0.073$). However, phenotypic effects were much weaker than those of the peak QTL markers (Fig. 2A and Table S8, B and C), which implies that these shared TE insertions are unlikely causal.

To support this conclusion, we crossed a new pair of WT and *ddm1-2* founder plants (Fig. 2B) [21] and selfed the F1 to select a single *DDM1*/*DDM1* F2 individual that was “epi-heterozygote” for all three QTL peak markers on chr 1, 2, and 4 [21]. After an additional selfing, we selected F3 progeny from the long and short extremes of the RL distribution. DNA methylation analysis confirmed the association of short and long primary roots with the *ddm1-2*–like and WT methylation states at the peak QTL marker, respectively (Fig. 2C) [21]. Furthermore, the tail-selected F3 individuals contained none of the shared TE insertions identified in the epiRIL population (Fig. 2D and Table S6). We thus conclude that the epiRIL QTL are most likely caused by the heritable (*ddm1-2* induced) loss of DNA methylation in the QTL intervals.

4.2.3 Candidate DMRs in the epiRIL QTL intervals

Next, we searched for putative causal DMRs in the RL and FT QTL intervals. We analyzed the methylomes (~165 base pair resolution) of the 123 epiRILs and their founders [19] and required candidate DMRs to be in approximate linkage disequilibrium with the peak QTL marker and displaying clear differences in DNA

methylation states and expression levels between the WT and the *ddm1-2* founder lines (Figs. S5 to S8). Our search revealed 325 candidate DMRs within the FT QTL

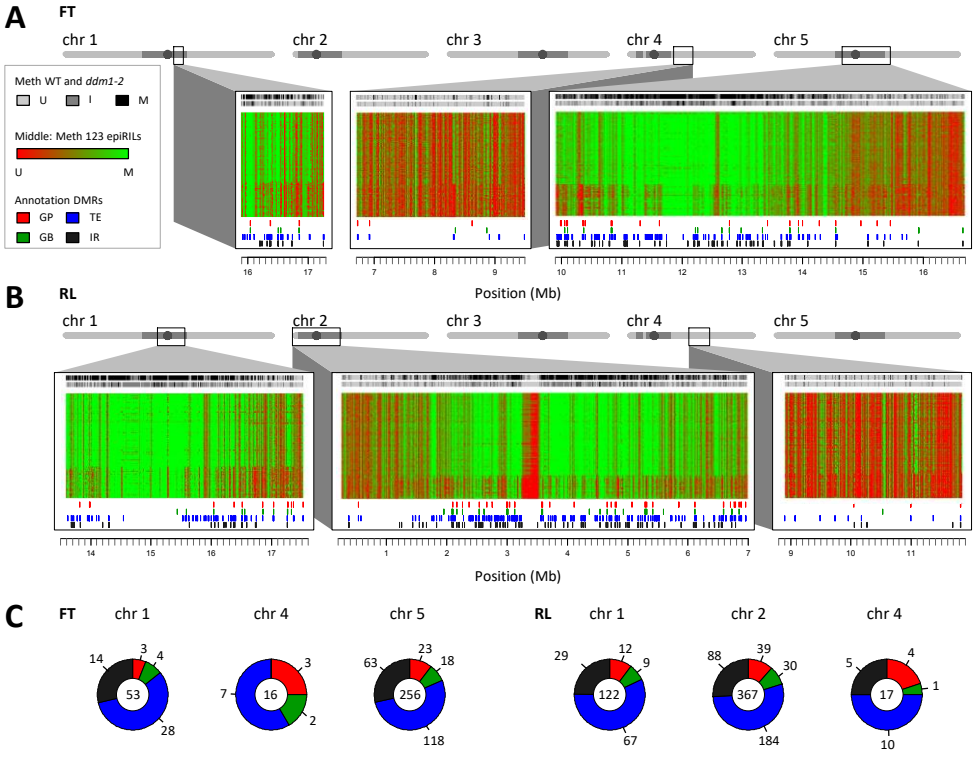

Figure 3. DNA methylation profiles of candidate DMRs in the epiRIL QTL intervals. (**A** and **B**) Location and annotation of candidate DMRs detected for FT (**A**) and RL (**B**). The top part of the rectangles shows the DNA methylation profile of the WT and *ddm1-2* parents, respectively (U, unmethylated; I, intermediate DNA methylation; M, high-level DNA methylation). The DNA methylation profiles of the epiRILs are indicated below and are ordered according to the epigenotype of the peak marker [from WT (top) to *ddm1-2* (bottom)]. The bottom part of the rectangles shows the annotations that overlap with the DMRs (GP, gene promoters; GB, gene bodies; TE, transposable element sequences; IR, intergenic regions). A schematic representation of each chromosome is plotted above each rectangle. (**C**) Number of candidate DMRs detected for each QTL interval (values inside circles) and the number of unique annotations (values outside circles) with which they overlap. DMRs can overlap multiple annotations (Tables S12 and S13). Colors and abbreviations are as in (**A**).

intervals (chr1, 53; chr4, 16; chr5, 256), which mapped to 44 unique genes (including promoter regions), 153 annotated TE sequences, and 77 intergenic regions (Fig. 3, A and C, and Tables S9, S10, and S12). For the RL QTL intervals, we detected 506 candidate DMRs (chr 1, 122; chr2, 367; chr4, 17), mapping to 71 unique genes (including promoter regions), 261 annotated TE sequences, and 122 intergenic regions (Fig. 3, B and C, and Tables S9, S11, and S13). Further analysis of these candidate DMRs did not identify any obvious flowering time and root length genes (Tables S10 and S11), which could be consistent with the lower amplitude of phenotypic variation observed among the epiRILs than among highly contrasted accessions [16]. However, we cannot rule out that the candidate DMRs are in LD with causal DMRs that could not be called with our method. Ultimately, fine-mapping approaches and targeted manipulation of selected DMRs will be required to identify causal regions.

4.3 Conclusion and discussion

Our analysis of the epiRILs demonstrates that induced DMRs can be stably inherited independently of DNA sequence changes and function as epigenetic quantitative trait loci (QTL^{epi}). Phenotypically, the detected QTL^{epi} have all the necessary properties to become targets of natural or artificial selection. Taking advantage of the single-nucleotide resolution methylomes of 138 natural accessions [4] (Table S14), we could show that about 30% of the heritable DMRs identified in the epiRIL population overlap with naturally occurring DMRs among these accessions (Figs. S9 and S10). Therefore, these epiRIL DMRs may have been historical targets of epimutations in the wild, either through trans-induced *ddm1*-like mutation events or else through still unknown mechanisms. This finding indicates in turn that DMRs could also act as QTL^{epi} in natural populations and thus constitute a measureable component of the so-called “missing heritability.” This possibility may have deep implications on how we delineate and interpret the heritable basis of complex traits.

Acknowledgments

We thank O. Bossdorf, C. Richards, T. Day, and K. Verhoeven for their input on an earlier version of this report. This work was supported by grants from the Netherlands Organization for Scientific Research (to F.J., R.C.J., R.W., and M.C.-T.); a University of Groningen Rosalind Franklin Fellowship to M.C.-T.; the Agence Nationale de la Recherche (ANR-09-BLAN-0237 EPIMOBILE to V.C. and P.W.; ANR-06-GPLA-010 TAG, Investissements d’Avenir ANR-10-LABX-54 MEMO LIFE, and ANR-

11-IDEX-0001-02 PSL* Research University to V.C.); and the European Union (EpiGeneSys FP7 Network of Excellence number 257082, to V.C.). S.C. and M.E. were supported by Ph.D. studentships from the Ministère de l'Enseignement Supérieur et de la Recherche, with additional support from the Fondation pour la Recherche Médicale (to S.C.). Sequence reads for the 52 epiRILs and parental lines are deposited at the European Bioinformatics Institute under accession number ERP004507.

Supplementary material

Supplementary text, figures and tables can be found at **http://www.sciencemag.org/content/343/6175/1145/suppl/DC1**.

References and notes

1. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**:204–220.
2. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci USA* **107**:8689–8694.
3. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**:916–919.
4. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, Ecker JR (2013) Patterns of population epigenomic diversity. *Nature* **495**:193–198.
5. Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, Waters AJ, Starr E, West PT, Tiffin P, Myers CL, Vaughn MW, Springer NM (2013) Epigenetic and genetic influences on DNA methylation variation in maize populations. *Plant Cell* **25**:2783–2797.
6. Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, Monk D, Hata K, Marques-Bonet T, Wang L, Esteller M (2013) DNA methylation contributes to natural human variation. *Genome Res* **23**:1363–1372.
7. Weigel D, Colot V (2012) Epialleles in plant evolution. *Genome Biol* **13**:249.
8. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**:245–249.

9. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334**:369–373.
10. Johannes F, Colot V, Jansen RC (2008) Epigenome dynamics: A quantitative genetics perspective. *Nat Rev Genet* **9**:883–890.
11. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, Bouchez D, Dillmann C, Guerche P, Hospital F, Colot V (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5**:e1000530.
12. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**:471–476.
13. Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T (2009) Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* **461**:423–426.
14. Kakutani T, Munakata K, Richards EJ, Hirochika H (1999) Meiotically and mitotically stable inheritance of DNA hypomethylation induced by *ddm1* mutation of *Arabidopsis thaliana*. *Genetics* **151**:831–838.
15. Teixeira FK, Heredia F, Sarazin A, Roudier F, Boccara M, Ciaudo C, Cruaud C, Poulain J, Berdasco M, Fraga MF, Voinnet O, Wincker P, Esteller M, Colot V (2009) A role for RNAi in the selective correction of DNA methylation defects. *Science* **323**:1600–1604.
16. Latzel V, Zhang Y, Karlsson Moritz K, Fischer M, Bossdorf O (2012) Epigenetic variation in plant responses to defense hormones. *Ann Bot* **110**:1423–1428.
17. Roux F, Colomé-Tatché M, Edelist C, Wardenaar R, Guerche P, Hospital F, Colot V, Jansen RC, Johannes F (2011) Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. *Genetics* **188**:1015–1017.
18. Johannes F, Colomé-Tatché M (2011) Quantitative epigenetics through epigenomic perturbation of isogenic lines. *Genetics* **188**:215–227.
19. Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E, Labadie K, Wincker P, Jacobsen SE, Jansen RC, Colot V, Johannes F (2012) Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci USA* **109**:16240–16245.

20. Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**:185–199.
21. Materials and methods and supplementary text are available as supporting material on Science Online.

Chapter 5

Epigenetic divergence is sufficient to trigger heterosis in *Arabidopsis thaliana*

Preprint:

Lauss K, Wardenaar R, van Hulten M, Guryev V, Keurentjes JJ, Stam M, Johannes F (2016) Epigenetic divergence is sufficient to trigger heterosis in *Arabidopsis thaliana*.

Chapter 5

Abstract

Despite the importance and wide exploitation of heterosis in commercial crop breeding, the molecular mechanisms behind this phenomenon are not well understood. Interestingly, there is growing evidence that beside genetic also epigenetic factors contribute to heterosis [1-3]. Here we used near-isogenic but epigenetically divergent parents to create epigenetic F1 hybrids (epiHybrids) in *Arabidopsis*, allowing us to quantify the contribution of epigenetics to heterosis. We measured traits such as leaf area (LA), growth rate (GR), flowering time (FT), main stem branching (MSB), rosette branching (RB), final plant height (HT) and seed yield (SY) and observed several strong positive and negative heterotic phenotypes among the epiHybrids. For LA and HT mainly positive heterosis was observed, while FT and MSB mostly displayed negative heterosis. Heterosis for FT, LA and HT could be associated with several heritable, differentially methylated regions (DMRs) in the parental genomes. These DMRs contain 35 (FT and LA) and 14 (HT) genes, which may underlie the heterotic phenotypes observed. In conclusion, our study indicates that epigenetic divergence can be sufficient to cause heterosis.

5.1 Introduction

Heterosis describes an F1 hybrid phenotype that is superior compared to the phenotype of its parent varieties. The phenomenon has been exploited extensively in agricultural breeding for decades and has improved crop performance tremendously [4, 5]. Despite its commercial impact, knowledge of the molecular basis underlying heterosis remains incomplete. Most studies mainly focused on finding genetic explanations, resulting in the classical dominance [4, 6, 7] and overdominance [7, 8] models describing heterosis. In line with genetic explanations it has been observed that interspecies hybrids often show a higher degree of heterosis than intraspecies hybrids, indicating that genetic distance correlates with the extent of heterosis [5, 9]. However, genetic explanations do often not sufficiently explain nor predict heterosis. There is growing evidence that also epigenetic divergence plays a role in heterosis [1-3]. It has for example been shown that altered epigenetic profiles at genes regulating circadian rhythm play an important role in heterotic *Arabidopsis* hybrids [10]. Moreover, heterotic hybrids of *Arabidopsis*, maize and tomato are shown to differ in levels of small regulatory RNAs and/or DNA methylation (5mC) relative to their parental lines [11–14]. Processes such as the transfer of 5mC between alleles (trans chromosomal methylation, TCM), or a loss of 5mC at one of the alleles (trans chromosomal demethylation, TCdM) have been

indicated to contribute to the observed remodeling of the epigenome [1, 13, 15]. Strikingly, some of these changes in 5mC levels have been shown to be stable over multiple generations [15, 16].

Hybrids are usually generated from parental lines that vary at both the genomic and epigenomic level and disentangling those two sources of variation is challenging. To overcome this limitation, we generated epigenetic *Arabidopsis thaliana* F1 hybrids (epiHybrids) from near-isogenic but epigenetically divergent parental lines by crossing Col-0 wild-type (Col-wt) as maternal parent to 19 near-isogenic *ddm1-2*-derived epigenetic recombinant inbred lines (epiRILs) [17] as the paternal parents (Fig. 1A). *DDM1* (*DECREASE IN DNA METHYLATION 1*) is a nucleosome remodeler and a *ddm1-2* deficiency leads to a severe loss of 5mC [18], primarily in long transposable elements and other repeat sequences [19]. EpiRILs carry chromosomes that are a mosaic of Col-wt and hypomethylated *ddm1-2*-derived genomic regions [17, 20, 21] (Fig. 1A). Nineteen epiRIL parental lines were selected that sample a broad range of 5mC divergence from the Col-wt reference methylome (Fig. 1B, Table S1). Besides, lines were chosen that have a wild-type methylation profile at *FWA* (Supplementary Fig. 1, Table S1), as loss of DNA methylation at the *FWA* (*FLOWERING WAGENINGEN*) locus is known to affect flowering time [22]. Furthermore, we selected for a range of phenotypic variation in two traits that have previously been monitored in the epiRILs, flowering time and root length (Table S1); outliers were excluded [21]. With our experimental design we could demonstrate, as proof-of-principle, the extent to which divergence in 5mC profiles in parental lines can contribute to heterosis.

5.2 Materials, methods and definitions

The phenotypic performance of the 19 epiHybrids and their parental lines was assessed by monitoring about 1090 plants (~28 replicates per line) for a range of quantitative traits: LA, GR, FT, MSB, RB, HT and SY (Tables S2-S7). The phenotypic observations for SY were inconsistent in a replication experiment, therefore those datasets were excluded from further analysis. The hybrids and parental lines were grown in parallel in a climate-controlled chamber with automated watering. The plants were randomized throughout the chamber to level out phenotypic effects caused by plant position. LA was measured up to 14 days after sowing (DAS), using an automated camera system (Fig. 1F), and growth rate (GR) was determined based on this data (SI text). FT was scored manually as opening of the first flower. After all plants started flowering, the plants were transferred to the greenhouse and grown to maturity. MSB, RB and HT were scored manually after harvesting of the plants.

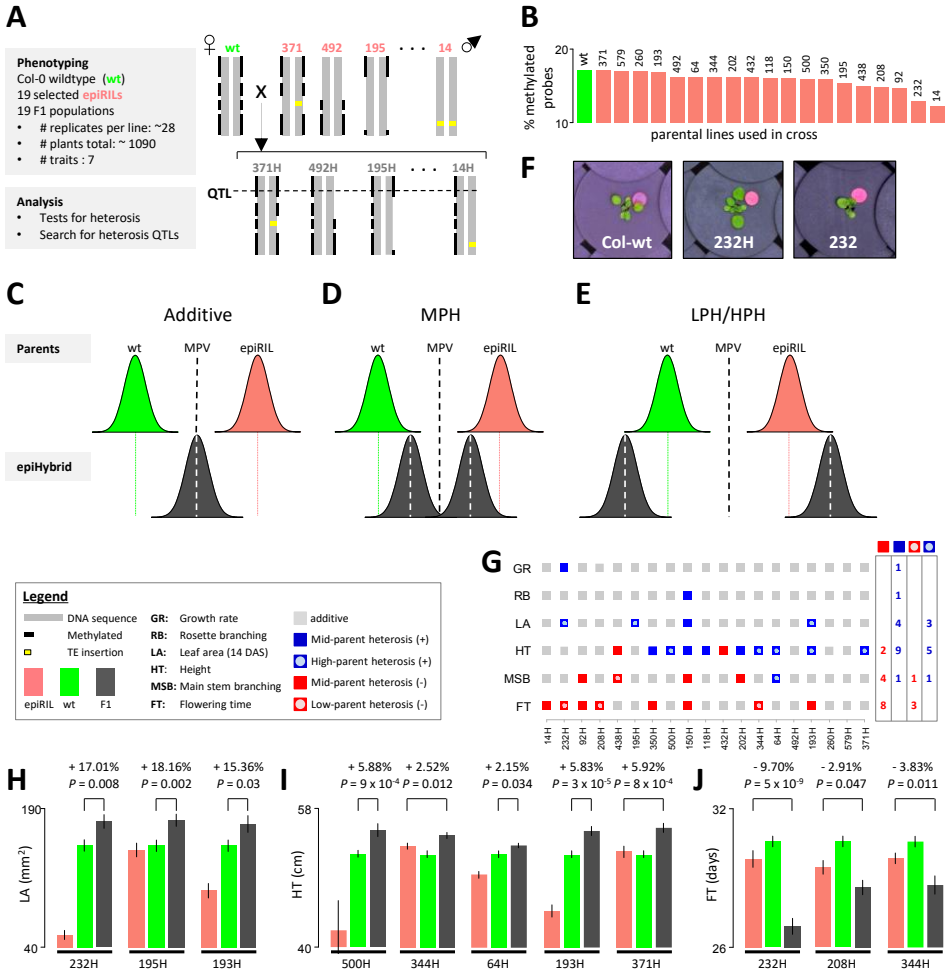

Figure 1. Heterosis occurs in epiHybrids. (**A**) Experimental setup. Lines are depicted schematically as one chromosome with the numbers indicating the epiRIL ID (e.g. 371 & 492) and the respective epiHybrid (e.g. 371H & 492H). (**B**) Genome-wide 5mC levels (*y*-axis) of the Col-wt line in green and the epiRIL parental lines in salmon. Numbers indicate the epiRIL IDs. The 5mC levels were calculated as the proportion of methylated MeDIP probes with respect to the total amount of probes. (**C - E**) Three classes of phenotypic effects monitored in the epiHybrids. The black dashed line indicates the mid-parent value. The green and salmon dashed lines indicate the mean performance of the parental lines. The white dashed lines indicate the mean performance of the epiHybrids. (**F**) Col-wt, epiHybrid 232H and epiRIL 232 at 13 days after sowing as an example for high-parent heterosis. (**G**) Phenotypic effects in six traits monitored across the 19 epiHybrids. The right panel summarizes positive and negative

Figure 1. (continued) … heterotic effects per trait. (**H - J**) Examples of epiHybrids exhibiting high-parent heterosis in leaf area and height (LA and HT; H and I), and low-parent heterosis in flowering time (FT; J) Error bars, ± 1 SEM. Deviation from high parent or low parent is shown in percent.

The extent of heterosis was evaluated by comparing the hybrid performance with its parental lines. We distinguished five effects (Fig. 1C-E): additivity, positive mid-parent heterosis (positive MPH), negative mid-parent heterosis (negative MPH), high-parent heterosis (HPH) and low-parent heterosis (LPH). An additive effect describes a hybrid performance that is equal or close to the average performance of the two parents (the mid-parent value, MPV). MPH refers to deviations in percent from the MPV in positive or negative direction. Hybrids displaying MPH are further tested for HPH and LPH, which describe hybrid performance exceeding the highest parent, or falling below the lowest parent, respectively. In crop breeding, the focus is usually on obtaining HPH and LPH as these present novel phenotypes that are outside the parental range. Depending on the trait monitored and commercial application, either HPH or LPH can be considered superior. For instance, early flowering may be preferable over late flowering; in such cases maximizing LPH may be desirable. For other traits, such as yield or biomass, it is more important to maximize HPH. However, in order to obtain a comprehensive view of hybrid performance it is informative to also track MPH in addition to LPH and HPH, because many mature traits may be affected by other traits that do not display fully penetrant heterotic effects.

5.3 Results

5.3.1 Observed heterotic phenotypes

We observed a remarkably wide range of heterotic phenotypes among the epiHybrids (Fig. 1G, Tables S2-19). The magnitude of these phenotypic effects was substantial (Fig. 1H-J, Supplementary Fig. 2, Tables S8-19) and similar to that typically seen in hybrids of *Arabidopsis* natural accessions [23, 24]. Many epiHybrids (16/19) exhibited significant MPH in at least one of the six monitored traits (FDR = 0.05, Fig. 1G). Across all hybrids and traits, we observed 30 cases of positive MPH and negative MPH. Among those, four cases show LPH and nine cases show HPH (Fig. 1G). Interestingly, in 11 out of the 17 cases of MPH the phenotypic means of the epiHybrids were in the direction of the phenotypic means of the epiRIL parent rather than in the direction of the Col-wt parent (Tables S2-7, F1 trend). Also all four LPH

and two of the HPH cases were in the direction of the epiRIL parent (Fig. 1I-J, Supplementary Fig. 2). This observation illustrates that *ddm1-2*-derived hypomethylated epialleles are often (partially) dominant over wild-type epialleles, which contrasts the situation seen in EMS screens where novel mutations typically act recessively.

We observed cases of HPH for LA, HT and MSB, and cases of LPH for FT and MSB. HPH for LA occurred in epiHybrids 232H, 195H and 193H (3/19 epiHybrids). Those epiHybrids significantly exceeded their best parent (Col-wt) by 17%, 18% and 15%, respectively (Fig. 1H, Table S19). Interestingly, although growth rate (GR) is developmentally related to LA, hybrid effects in GR were only moderately, albeit positively, correlated with LA (rho = 0.57, P = 0.02), which implies that LA heterosis is determined by other traits besides GR. For HT we detected five cases of significant HPH with up to 6% increases in HT (Fig. 1I, Table S14). One may expect LA HPH to strongly correlate with HT HPH, as the rosette is providing nutrients for the developing shoot [25]. However, HPH for both LA and HT occurred only in one epiHybrid (193H; Fig. 1G). For MSB, we detected one case of HPH (64H; Fig. 1G and Supplementary Fig. 2).

Besides positive heterosis, our phenotypic screen revealed strong negative heterotic effects for FT (earlier flowering) and MSB (less main stem branching). Significant LPH occurred in the epiHybrids 232H, 208H and 344H (FT) and 438H (MSB) (Fig. 1J, Supplementary Fig. 2, Tables S15, S17). In the most prominent case for FT (232H), FT was about 10% earlier than that of the earliest flowering parent. 208H and 344H flowered 3% and 4% earlier than their lowest parent (epiRIL 208 and epiRIL 344), respectively. 438H showed 14% less MSB than the lowest parent (Supplementary Fig. 2).

5.3.2 Confirmation with replicate experiments

The reproducibility of our findings was tested by performing replicate experiments, using seeds from newly performed crosses and the same climate controlled growth chamber as before. We focused on epiHybrids that exhibited relatively strong positive or negative heterotic phenotypes in the initial screen (193H, 150H, 232H; Fig. 1G), and measured LA, FT and HT. We found that the direction of the heterotic effects in LA, FT and HT was reproducible in all cases tested (Fig. 2A and B). Importantly, the LA and HT HPH observed for 193H, and the strong FT LPH for 232H were perfectly reproducible, while LA HPH observed for 232H became positive MPH (Fig. 2A). Taken together, these results show that the heterotic effects observed in the epiHybrids are relatively stable for LA, HT and FT, even across fresh parental seed

batches and independently performed crosses, which is not always the case for *Arabidopsis* phenotypes [26].

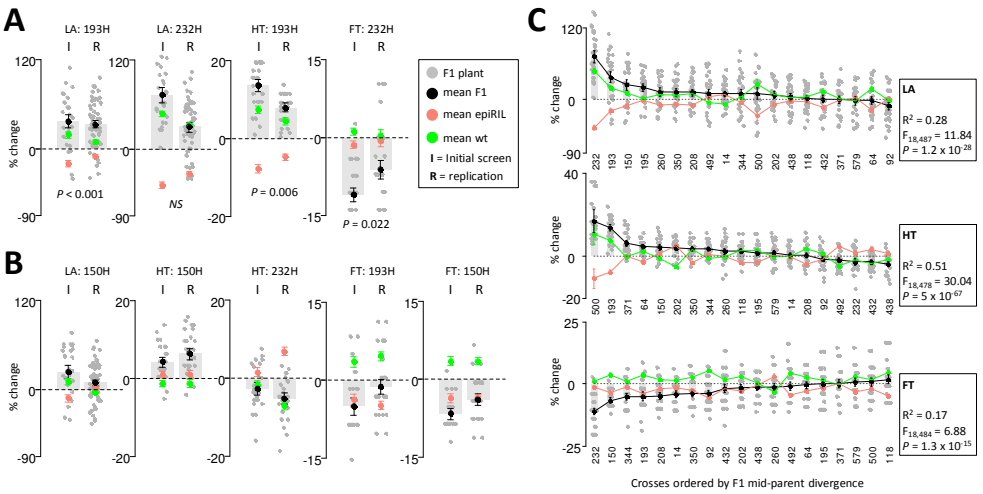

Figure 2. Confirmation of mid-parent (MP) divergence in the initial screen and replicate experiment for epiHybrids 150H, 193H and 232H. (**A**) Results for cases of HPH and LPH for LA, HT and FT in initial experiment. (**B**) Results for traits showing less eminent phenotypic effects for LA, HT and FT. The mid-parent value (MPV) is shown as a dashed horizontal line and the MP divergence is shown as change from MPV in percent. To illustrate the F1 epiHybrid distribution for each trait, the individual replicate plants are depicted as dots. (**C**) F1 MP divergence for LA, HT and FT for all epiHybrids. The MPV is shown as a horizontal dashed line and MP divergence is shown as change from MPV in percent. The epiHybrids are ordered from highest (left) to lowest (right) F1 MP divergence. To illustrate the F1 epiHybrid distribution for each trait, the individual replicate plants are depicted as dots. Variance component analysis was used to estimate how much of the total variation in MP divergence can be explained by between-cross variation. The F-statistic from this analysis is shown in the boxes.

4.1.1 Interval mapping using mid-parent divergence as phenotype

To understand the sources of the LA, HT and FT heterotic effects observed among the ~530 epiHybrid plants, we calculated the phenotypic divergence of each epiHybrid plant from its respective mid-parent value. Using variance component

analysis we estimated that 17%, 28% and 51% of the total variation in mid-parent divergence for FT, LA and HT, respectively, can be attributed to (epi)genomic differences between the Col-wt and epiRILs used for the crosses (Fig. 2C, Table S20, SI text). Global 5mC divergence between the Col-wt and the epiRIL parental lines could not account for this variation (Supplementary Fig. 3). We therefore reasoned that heterotic phenotypes are due to (partial) dominance effects caused by specific regions being epi-heterozygous for an epiRIL-inherited hypomethylated epiallele (*U*) and a Col-wt-inherited methylated epiallele (*M*). To test this possibility, we used the methylomes of Col-wt and the epiRIL parents [20] to predict epi-homozygous (*MM*) and epi-heterozygous (*MU*) regions in the genomes of the epiHybrids (Fig. 3A, SI text), and assessed whether heritable epigenetic differences at specific loci could explain the variation in MPH among crosses (Supplementary Fig. 4).

The analysis revealed two QTLs on chromosome (chr) 3 contributing to the between-cross variation in MPH in FT (QTL 1: LOD=3.12, 37.62 cM; QTL 2: LOD=3.33, 101.44 cM, Fig. 3B; Table S21). EpiHybrids epi-heterozygous (*MU*) at these loci showed significant negative MPH compared to their epi-homozygous (*MM*) counterparts (Fig. 3C). While not significant at the genome-wide scale (Fig. 3B), the same two QTLs had substantial suggestive effects on LA heterosis in the opposite direction than FT (Fig. 3B and C), indicating that both QTLs act pleiotropically.

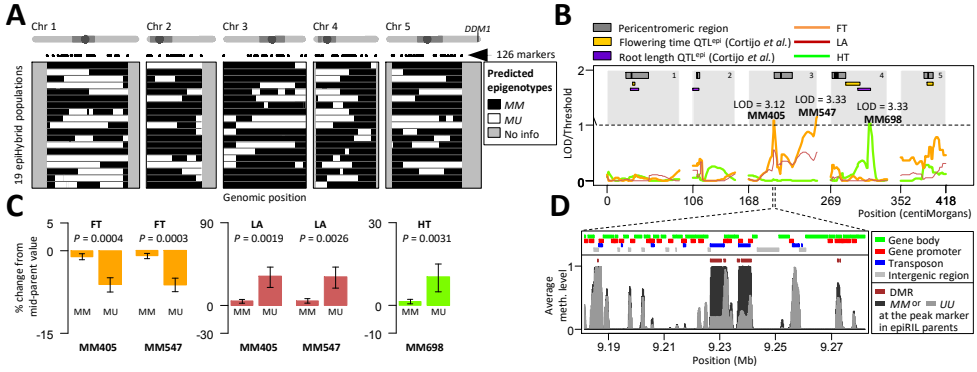

Figure 3. Interval mapping approach detects significant QTLs for mid-parent divergence. (**A**) Genome-wide patterns of Col-wt and *ddm1-2* inherited epi-haplotypes in the (epi)genomes of the parental epiRILs used in this study. (**B**) QTL profiles for FT, HT and LA. Published QTLs^{epi} for root length and flowering time are shown. (**C**) Effect direction of the QTLs. Error bars, ± 1 SE of the Estimate (SEE). (**D**) Zoom in of one of the QTL intervals of FT. The top panel shows the annotations along the genome. The bottom panel shows the locations of candidate DMRs and the average methylation level along the genome for epiRIL parents that are either methylated (*MM*) or unmethylated (*UU*) at the peak marker.

We also detected a single QTL locus on chr 4 (LOD = 3.33, 56.00 cM) that contributes to the between-cross variation in MPH for HT (Fig. 3B, Table S21). In this case, *MU* epiHybrids showed significant positive MPH compared to *MM* epiHybrids (Fig. 3C). Interestingly, the HT QTL overlaps with a previously identified QTL^{epi} for root length in the epiRILs [21]. The same study identified $QTLs^{epi}$ associated with FT [21] that we did not detect here (Fig. 3B), implying that different regions may play a role in FT trait variation than in FT heterosis.

5.3.4 Potential causal variants in the epiHybrid QTL intervals

The detection of heterosis QTLs for FT, LA and HT provided a rationale to search for causal variants in the QTL confidence intervals. TE-associated structural variants (TEASVs) are known to occur at low frequency in a *ddm1-2*-derived DNA hypomethylated background [17, 21, 27, 28], hence we re-analyzed whole-genome sequencing data from the epiRIL parents [21] for TEASVs but did not detect any that could account for the QTL effects, suggesting that the QTLs most likely have an epigenetic basis (SI text). Indeed, a thorough analysis of the methylomes of the parental epiRILs, using the available MeDIP tiling array data [20], identified 55 and 18 potentially causal differentially methylated regions (DMRs) in the FT, LA and HT QTL regions, mapping to 35 and 14 unique genes, respectively (Fig. 3D, Supplementary Fig. 5–9, Tables S22-S26, SI text). Potentially interesting genes in the candidate regions of the FT/LA QTLs (Table S25) include for example RPL5A, which was shown to affect development through regulating auxin and influencing leaf shape and patterning [29, 30], and AT3G26480, a protein that shows partial homology to GTS1, which has been implemented in biomass accumulation [31]. Another potentially interesting candidate is Chup1, which is crucial for chloroplast movement in leaves in response to light [32]. These candidate genes provide excellent targets for follow-up studies.

5.4 Discussion

In a recently published study, heterosis for rosette area was reported in an epigenetic F1 hybrid generated by crossing a *met1*-derived epiRIL with Col-wt [3]. *DNA-METHYLTRANSFERASE1* (*MET1*) is involved in maintenance of DNA methylation at cytosines in CG sequence context and a mutation in this gene causes a severe loss of DNA methylation in the CG and CHH context [33]. Heterosis was observed in a parent-of-origin manner; the reciprocal cross did not result in heterosis [3]. This suggests that the heterosis detected may be due to an effect of the maternal

cytoplasm rather than differences in epigenetic marks in the parental genomes. Here, we used Col-wt as maternal parent in all crosses to specifically monitor phenotypic effects associated with the epiRIL methylomes. We observed a wide range of heterotic effects, and our proof-of-principle QTL mapping approach indicated that these phenotypic effects are very likely attributable to methylation differences between Col-wt and the epiRILs. Moreover, our results, together with those of Dapp *et al.* [3], indicate that heterosis in F1 hybrids generated from epigenetically divergent lines may be a more general phenomenon. It will be interesting to see whether these findings have implications for future crop breeding.

Acknowledgments

We thank F. Becker, I. Hövel, D. Angorro, R. Kooke, J.A. Bac-Molenaar, M. Tark-Dame, P. Sanderson, M. Koini, T. Bey, B. Weber, L. Tikovsky and Unifarm Wageningen for technical support during sowing or phenotyping. We thank H. Westerhoff for discussion and critically reading the manuscript. The phenotyping was supported by D. Vreugdenhil and funded by an ETP Hotelproject Grant. K. Lauss was supported by the Centre for Improving Plant Yield (CIPY; part of the Netherlands Genomics Initiative and the Netherlands Organization for Scientific Research). F. Johannes was supported by the Technische Universität München – Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Program under grant agreement n° 291763.

Contributions

K.L., M.S. and F.J. designed the study, interpreted the data and wrote the manuscript with contributions from J.J.B.K. and R.W. K.L. and M.H.A.v.H. planned and performed the phenotypic screen. F.J. and R.W. performed the data analysis. V.G. analyzed sequencing data of the epiRIL parents.

Supplementary material

Supplementary text, figures and tables can be found at **www.johanneslab.org/internal**. You can get access to the files with the following username and password:

Username: **guest@johanneslab.org**
Password: guestjlab16

References

1. Groszmann M, Greaves IK, Fujimoto R, Peacock WJ, Dennis ES (2013) The role of epigenetics in hybrid vigour. *Trends Genet* **29**:684-690.
2. Springer NM (2013) Epigenetics and crop improvement. *Trends Genet* **29**:241-247.
3. Dapp M, Reinders J, Bédiée A, Balsera C, Bucher E, Theiler G, Granier C, Paszkowski J (2015) Heterosis and inbreeding depression of epigenetic *Arabidopsis* hybrids. *Nat Plants* **1**:15092.
4. Schnable PS, Springer NM (2013) Progress toward understanding heterosis in crop plants. *Annu Rev Plant Biol* **64**:71-88.
5. Chen ZJ (2010) Molecular mechanisms of polyploidy and hybrid vigor. *Trends Plant Sci* **15**:57-71.
6. Jones DF (1917) Dominance of linked factors as a means of accounting for heterosis. *Genetics* **2**:466-479.
7. Crow JF (1998) 90 years ago: the beginning of hybrid maize. *Genetics* **148**:923-928.
8. Crow JF (1948) Alternative Hypotheses of Hybrid Vigor. *Genetics* **33**:477-487.
9. East EM (1936) Heterosis. *Genetics* **21**: 375-397.
10. Ni Z, Kim ED, Ha M, Lackey E, Liu J, Zhang Y, Sun Q, Chen ZJ (2009) Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature* **457**:327-331.
11. Groszmann M, Greaves IK, Albertyn ZI, Scofield GN, Peacock WJ, Dennis ES (2011) Changes in 24-nt siRNA levels in *Arabidopsis* hybrids suggest an epigenetic contribution to hybrid vigor. *Proc Natl Acad Sci USA* **108**:2617-2622.
12. Barber WT, Zhang W, Win H, Varala KK, Dorweiler JE, Hudson ME, Moose SP (2012) Repeat associated small RNAs vary among parents and following hybridization in maize. *Proc Natl Acad Sci USA* **109**:10444-10449.
13. Shivaprasad PV, Dunn RM, Santos BA, Bassett A, Baulcombe DC (2012) Extraordinary transgressive phenotypes of hybrid tomato are influenced by epigenetics and small silencing RNAs. *EMBO J* **31**:257-266.
14. Shen H, He H, Li J, Chen W, Wang X, Guo L, Peng Z, He G, Zhong S, Qi Y, Terzaghi W, Deng XW (2012) Genome-wide analysis of DNA methylation and gene expression changes in two *Arabidopsis* ecotypes and their reciprocal hybrids. *Plant Cell* **24**:875-892.

15. Greaves IK, Groszmann M, Ying H, Taylor JM, Peacock WJ, Dennis ES (2012) Trans chromosomal methylation in *Arabidopsis* hybrids. *Proc Natl Acad Sci USA* **109**:3570-3575.
16. Greaves IK, Groszmann M, Wang A, Peacock WJ, Dennis ES (2014) Inheritance of trans chromosomal methylation patterns from *Arabidopsis* F1 hybrids. *Proc Natl Acad Sci USA* **111**:2017-2022.
17. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, Bouchez D, Dillmann C, Guerche P, Hospital F, Colot V (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5**:e1000530.
18. Vongs A, Kakutani T, Martienssen RA, Richards EJ (1993) *Arabidopsis thaliana* DNA methylation mutants. *Science* **260**:1926–1928.
19. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D (2013) The *Arabidopsis* nucleosome remodeler *DDM1* allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **153**:193-205.
20. Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E, Labadie K, Wincker P, Jacobsen SE, Jansen RC, Colot V, Johannes F (2012) Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci USA* **109**:16240–16245.
21. Cortijo S, Wardenaar R, Colomé-Tatché M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury JM, Wincker P, Roudier F, Jansen RC, Colot V, Johannes F (2014) Mapping the epigenetic basis of complex traits. *Science* **343**:1145-1148.
22. Soppe WJ, Jacobsen SE, Alonso-Blanco C, Jackson JP, Kakutani T, Koornneef M, Peeters AJ (2000) The late flowering phenotype of *fwa* mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Mol Cell* **6**:791-802.
23. Groszmann M, Gonzalez-Bayon R, Greaves IK, Wang L, Huen AK, Peacock WJ, Dennis ES (2014) Intraspecific *Arabidopsis* hybrids show different patterns of heterosis despite the close relatedness of the parental genomes. *Plant Physiol* **166**:265-280.
24. Wang L, Greaves IK, Groszmann M, Wu LM, Dennis ES, Peacock WJ (2015) Hybrid mimics and hybrid vigor in *Arabidopsis*. *Proc Natl Acad Sci USA* **112**:E4959-E4967.

25. Bennett E, Roberts JA, Wagstaff C (2012) Manipulating resource allocation in plants. *J Exp Bot* **63**:3391-3400.
26. Massonnet C, Vile D, Fabre J, Hannah MA, Caldana C, Lisec J, Beemster GT, Meyer RC, Messerli G, Gronlund JT, Perkovic J, Wigmore E, May S, Bevan MW, Meyer C, Rubio-Díaz S, Weigel D, Micol JL, Buchanan-Wollaston V, Fiorani F, Walsh S, Rinn B, Gruissem W, Hilson P, Hennig L, Willmitzer L, Granier C (2010) Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three *Arabidopsis* accessions cultivated in ten laboratories. *Plant Physiol* **152**:2142-2157.
27. Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T (2009) Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* **461**:423-426.
28. Kooke R, Johannes F, Wardenaar R, Becker F, Etcheverry M, Colot V, Vreugdenhil D, Keurentjes JJ (2015) Epigenetic basis of morphological variation and phenotypic plasticity in *Arabidopsis thaliana*. *Plant Cell* **27**:337-348.
29. Rosado A, Li R, van de Ven W, Hsu E, Raikhel NV (2012) *Arabidopsis* ribosomal proteins control developmental programs through translational regulation of auxin response factors. *Proc Natl Acad Sci USA* **109**:19537-19544.
30. Pinon V, Etchells JP, Rossignol P, Collier SA, Arroyo JM, Martienssen RA, Byrne ME (2008) Three *PIGGYBACK* genes that specifically influence leaf patterning encode ribosomal proteins. *Development* **135**:1315-1324.
31. Gachomo EW, Jimenez-Lopez JC, Baptiste LJ, Kotchoni SO (2014) GIGANTUS1 (GTS1), a member of Transducin/WD40 protein superfamily, controls seed germination, growth and biomass accumulation through ribosome-biogenesis protein interactions in *Arabidopsis thaliana*. *BMC Plant Biol* **14**:37.
32. Oikawa K, Kasahara M, Kiyosue T, Kagawa T, Suetsugu N, Takahashi F, Kanegae T, Niwa Y, Kadota A, Wada M (2003) Chloroplast unusual positioning 1 is essential for proper chloroplast positioning. *Plant Cell* **15**:2805-2815.
33. Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE (2014) Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nat Struct Mol Biol* **21**:64-72.

Chapter 6

Rate, spectrum, and evolutionary dynamics of spontaneous epimutations

---

Published as:

van der Graaf A*, Wardenaar R*, Neumann DA, Taudt A, Shaw RG, Jansen RC, Schmitz RJ, Colomé-Tatché M, Johannes F (2015) Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proc Natl Acad Sci USA* **112**:6676-6681

*Equal contribution

Chapter 6

Abstract

Stochastic changes in cytosine methylation are a source of heritable epigenetic and phenotypic diversity in plants. Using the model plant *Arabidopsis thaliana*, we derive robust estimates of the rate at which methylation is spontaneously gained (forward epimutation) or lost (backward epimutation) at individual cytosines and construct a comprehensive picture of the epimutation landscape in this species. We demonstrate that the dynamic interplay between forward and backward epimutations is modulated by genomic context and show that subtle contextual differences have profoundly shaped patterns of methylation diversity in *Arabidopsis thaliana* natural populations over evolutionary timescales. Theoretical arguments indicate that the epimutation rates reported here are high enough to rapidly uncouple genetic from epigenetic variation, but low enough for new epialleles to sustain long-term selection responses. Our results provide new insights into methylome evolution and its population-level consequences.

Significance

Changes in the methylation status of cytosine nucleotides are a source of heritable epigenetic and phenotypic diversity in plants. Here we derive robust estimates of the rate at which cytosine methylation is spontaneously gained (forward epimutation) or lost (backward epimutation) in the genome of the model plant *Arabidopsis thaliana*. We show that the forward–backward dynamics of selectively neutral epimutations have a major impact on methylome evolution and shape genome-wide patterns of methylation diversity among natural populations in this species. The epimutation rates presented here can serve as reference values in future empirical and theoretical population epigenetic studies in plants.

6.1 Introduction

Plant genomes make extensive use of cytosine methylation to control the expression of transposable elements (TEs) and genes [1]. Despite its tight regulation, methylation losses or gains at individual cytosines or clusters of cytosines can emerge spontaneously, in an event termed “epimutation” [2, 3]. Many examples of segregating epimutations have been documented in experimental and wild populations of plants and in some cases contribute to heritable variation in phenotypes independently of DNA sequence variation [4, 5]. These observations have led to much speculation about the role of DNA methylation in plant evolution

[6–8], and its potential in breeding programs [9]. In the model plant *Arabidopsis thaliana*, spontaneous methylation changes at CG dinucleotides accumulate in a rapid but nonlinear fashion over generations [2, 3, 10], thus pointing to high forward–backward epimutation rates [11]. Precise estimates of these rates are necessary to be able to quantify the long-term dynamics of epigenetic variation under laboratory or natural conditions, and to understand the molecular mechanisms that drive methylome evolution [12–14]. Here we combine theoretical modeling with high-resolution methylome analysis of multiple independent *Arabidopsis thaliana* mutation accumulation (MA) lines [15], including measurements of methylation changes in continuous generations, to obtain robust estimates of forward and backward epimutation rates.

6.2 Results

We joined whole-genome MethylC-seq [16] data from two earlier MA studies [2, 3] with extensive multigenerational MethylC-seq measurements from three additional MA lines (Fig. 1A and SI Appendix, Tables S1–S6). The first of these new MA lines (MA1_3) was propagated for 30 generations and includes measurements for 13 (nearly) consecutive generations (Fig. 1A). The other two MA lines (MA2_3) were propagated for 17 generations and were measured every four generations on average (Fig. 1A). These new data therefore allowed us to track epimutation dynamics over a large number of generations and at high temporal resolution. We constructed base pair-resolution methylation maps for all sequenced individuals (SI Appendix). To obtain a measure of genome-wide methylation divergence between any two individuals in a given MA pedigree, we calculated the proportion of differentially methylated cytosines in sequence contexts CG, CHG, and CHH (where H can be any base but G). For these calculations we used a set of consensus cytosines for which all individuals in the pedigrees had coverage of more than three reads (SI Appendix). This read coverage cutoff was found to be sufficient for robust downstream analyses (SI Appendix, Figs. S1 and S2). Consistent with previous reports [2, 3, 10], genome-wide methylation divergence at CG dinucleotides increased with divergence time in all pedigrees (Fig. 1B), but not in sequence contexts CHG and CHH (SI Appendix, Fig. S3). This distinction reflects intrinsic differences in the maintenance pathways that target these three contexts [1] and possibly also increased measurement error and cellular heterogeneity for non-CG methylation (SI Appendix, Fig. S4).

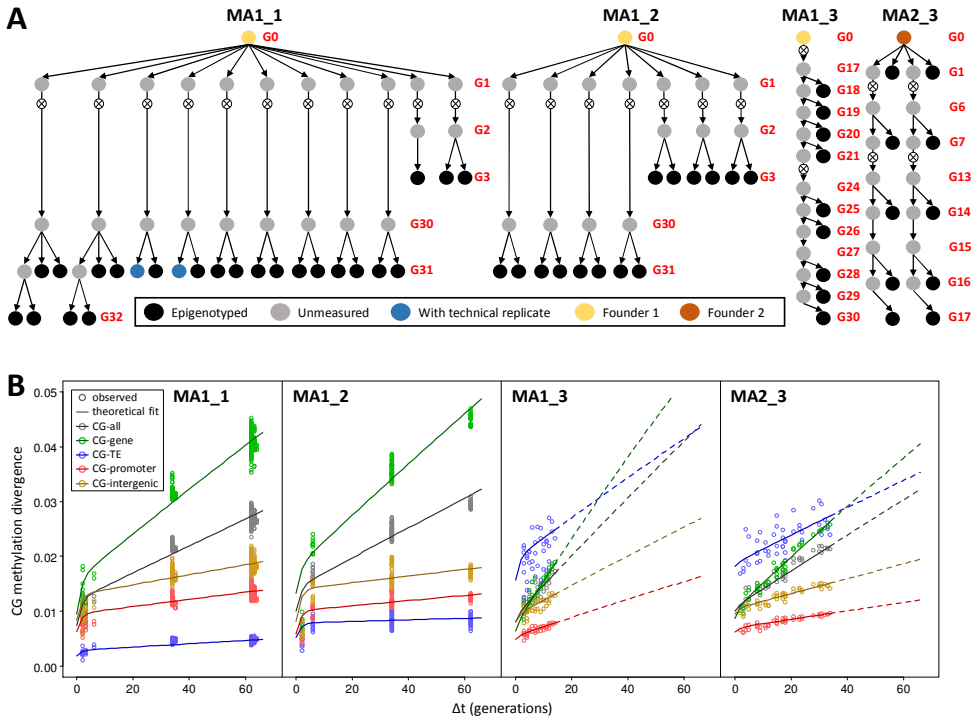

Figure 1. (**A**) Overview of pedigrees of mutation accumulation lines (MA lines). Red numbers indicate the number of generations from common founder. The MA1_1 and MA1_2 lines were originally created by Shaw *et al.* [15] and their methylomes were presented in Becker *et al.* [2] and Schmitz *et al.* [3], respectively. The MA1_3 and MA2_3 were generated in this study. (**B**) Measured CG methylation divergence (circles) with corresponding theoretical fits (lines) as a function of divergence time between two individuals in a given pedigree (Δt = total divergence time in generations between any two individuals). Dashed lines indicate extrapolation from the fitted model. Divergence values for CG-TE in datasets MA1_3 and MA2_3 were susceptible to low sequencing depth in these experiments and showed increased measurement noise (SI Appendix). Model-based analysis of all MA pedigrees revealed that the highly nonlinear divergence until generation eight is due to the fixation of segregating epi-heterozygote founder loci (SI Appendix), rather than the result of recurrent cycles of forward and backward epimutations as previously suggested [11].

6.2.1 Neutral epimutation model

To quantify CG methylation divergence in the MA lines as a function of divergence time (measured in generations) and forward–backward epimutation rates, we developed a theoretical model similar to those used in the analysis of regular systems of inbreeding (Materials and methods and SI Appendix). Briefly, the model assumes that an unmethylated cytosine (c^u) can become methylated (c^m) with probability α and likewise a methylated cytosine can become unmethylated with probability β. We arbitrarily define α as the forward and β as the backward epimutation rate per generation per haploid methylome. Transitions of diploid epigenotypes ($c^m c^m$, $c^u c^u$, or $c^m c^u$) from one generation to the next are modeled through a transition matrix where the elements of the matrix are determined by the Mendelian segregation of epialleles (c^m or c^u) and the rates α and β. Consistent with the MA experimental design, selection on epigenotypes during inbreeding is assumed to be absent, so that the cumulative divergence among lines is driven solely by neutral epigenetic drift. Estimates for the unknown epimutation rates are obtained by fitting our model to the CG methylation divergence data of each MA pedigree separately (Materials and methods).

6.2.2 Estimates of global CG epimutation rates

As shown in Figure 1B, our model provides an excellent fit to the data, which suggests that the observed divergence patterns among the MA lines are largely the result of the transgenerational accumulation of selectively neutral epimutations. Model-based estimates for the forward and backward CG epimutation rates (CG-all) were $2.56 \cdot 10^{-4}$ and $6.30 \cdot 10^{-4}$, respectively (Table 1 and SI Appendix, Table S7). These estimates are similar to the value provided by Schmitz *et al.* [3] ($4.46 \cdot 10^{-4}$) but illustrate that methylation loss at CG dinucleotides is globally three times as likely as methylation gain. The ratio of loss to gain (β/α), also known as the mutational bias parameter, is an important quantity: It determines the CG methylation content of the *Arabidopsis thaliana* genome over evolutionary timescales. Assuming that the *Arabidopsis thaliana* methylome is at equilibrium, the estimated CG forward-backward epimutation rates imply that -in the absence of selection or gene conversion- about 30 % of all CG sites should be methylated and about 70 % are unmethylated, which is consistent with actual measurements [16, 17].

Table 1. Estimates of forward and backward epimutation rates.

Context	α	Range (α)		β	Range (β)		β/α	Range (β/α)	
CG-all	$2.56 \cdot 10^{-4}$	$2.08 \cdot 10^{-4}$	$3.69 \cdot 10^{-4}$	$6.30 \cdot 10^{-4}$	$3.23 \cdot 10^{-4}$	$1.13 \cdot 10^{-3}$	2.36	1.55	3.24
CG-gene	$3.48 \cdot 10^{-4}$	$2.77 \cdot 10^{-4}$	$4.87 \cdot 10^{-4}$	$1.47 \cdot 10^{-3}$	$9.46 \cdot 10^{-4}$	$2.45 \cdot 10^{-3}$	4.24	2.84	5.10
CG-TE*	$3.24 \cdot 10^{-4}$	$1.68 \cdot 10^{-4}$	$4.80 \cdot 10^{-4}$	$1.20 \cdot 10^{-5}$	$7.76 \cdot 10^{-6}$	$1.62 \cdot 10^{-5}$	0.040	0.034	0.046
CG-promoter	$5.17 \cdot 10^{-5}$	$2.92 \cdot 10^{-5}$	$9.33 \cdot 10^{-5}$	$5.88 \cdot 10^{-4}$	$1.33 \cdot 10^{-4}$	$1.40 \cdot 10^{-3}$	11.4	4.16	15.08
CG-intergenic	$1.15 \cdot 10^{-4}$	$6.13 \cdot 10^{-5}$	$1.70 \cdot 10^{-4}$	$3.25 \cdot 10^{-4}$	$6.36 \cdot 10^{-5}$	$7.69 \cdot 10^{-4}$	2.83	0.47	4.80

We assume that an unmethylated cytosine (c^u) can become methylated (c^m) with probability α, and likewise a methylated cytosine can become unmethylated with probability β. We arbitrarily define α as the forward and β as the backward epimutation rate per generation per haploid methylome. Shown are model-based estimates for α and β as an average of the MA1_1, MA1_2, MA1_3, and MA2_3 datasets, as well as the range of these estimates across datasets (range). The asterisk indicates that the average estimate was based only on the MA1_1 and the MA1_2 data (SI Appendix). These estimates can be considered robust, because the different MA pedigrees varied considerably in terms of plant material, growth conditions, and sequencing approach (SI Appendix, Table S1).

6.2.3 Estimates of annotation-specific CG epimutation rates

We examined the extent to which CG epimutation rates depend on genomic context. To do this, we separated all CGs according to annotation (gene bodies, promoters, TEs, and intergenic regions; *see* SI Appendix). Although annotation-specific CG epimutation rates were approximately within the same order of magnitude (Table 1 and SI Appendix, Table S7), subtle differences in these rates had a substantial impact on differential divergence of CG methylation across annotation categories (Fig. 1B). The highest combined forward and backward rates were found for CGs in gene bodies (CG-gene), which were $3.48 \cdot 10^{-4}$ and $1.47 \cdot 10^{-3}$, respectively (Table 1 and SI Appendix, Table S7). By contrast, the lowest rates were found for CGs in TEs (CG-TEs, forward: $3.24 \cdot 10^{-4}$ and backward: $1.20 \cdot 10^{-5}$). As a result of these low epimutation rates, methylation divergence for CG-TEs was much less pronounced (Fig. 1B), resembling the divergence patterns seen for CHG and CHH contexts (SI Appendix, Fig. S3). This observation suggests that CG-TEs come under the influence of silencing pathways that primarily target neighboring CHHs and CHGs [18–20]. Indeed, CG-TE was the only annotation category in which the ratio of backward to forward epimutation rates was less than unity (Table 1 and SI Appendix, Table S7), which implies that gain of methylation is strongly favored over methylation loss.

6.2.4 Genome architecture and chromatin environment predict CG methylation divergence patterns along chromosomes

Because CG epimutation rates are annotation-specific, we predicted that methylation divergence closely tracks annotation density along chromosomes. To test this, we moved in a 1-Mb sliding window along the genome (step size 100 kb) and calculated the divergence between MA lines as expected from our model after 31 generations of independent selfing (Fig. 2B and SI Appendix). Our calculations predicted that CG-methylation divergence is low in TE-rich pericentromeric regions and high in gene-rich chromosome arms (Fig. 2B and SI Appendix, Figs. S5 and S6). Remarkably, these predictions strongly agreed with the observed divergence patterns at the genome-wide scale (R^2 = 0.74, P < 0.0001).

An alternative, or complementary, explanation is that the annotation-specific divergence patterns are simply a reflection of the genome-wide distribution of heterochromatic domains, which would explain the clear partitioning between pericentromeres and euchromatin. To test this directly, we reanalyzed recent ChIP-seq data on histone variant H2A.W [21], a proxy for heterochromatin, and estimated epimutation rates for CGs in regions that were either enriched or depleted for H2A.W (SI Appendix). We used these rates in combination with the genome-wide density distribution of H2A.W to derive predictions of CG-methylation divergence patterns. Our analysis revealed that, at the genome-wide scale, heterochromatin-based predictions were approximately equivalent to annotation-based predictions (R^2 = 0.72, P < 0.0001, SI Appendix, Fig. S5), suggesting that chromatin environment is a sufficient and parsimonious explanation for the observed divergence patterns along chromosomes. These results further indicate that the maintenance of methylation at CG dinucleotides is slightly more error-prone in regions of open chromatin compared with more compact regions, probably as a by-product of active transcription.

6.2.5 The spectrum of neutral epimutations shapes CG methylation diversity in natural populations

An intriguing question is to what extent the epimutation landscape in the MA lines provides insights into the mechanisms that shape CG methylation diversity in *Arabidopsis thaliana* natural populations, which are the outcome of long and complex evolutionary processes. To assess this we reanalyzed MethylC-seq data from a large number of accessions collected from across the Northern Hemisphere

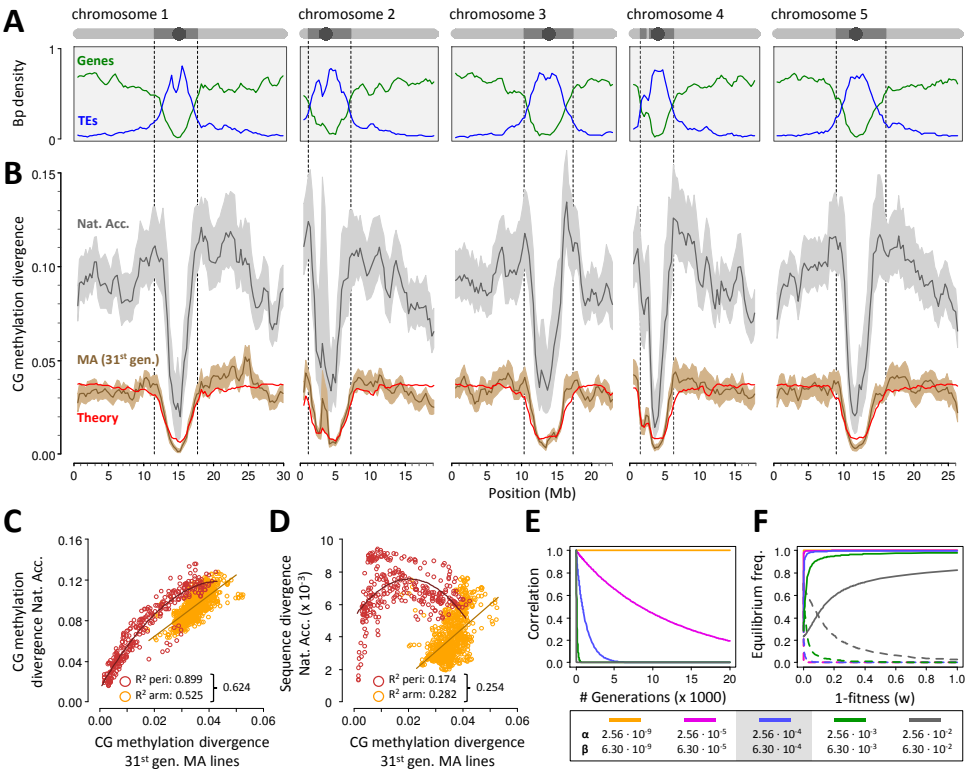

Figure 2. (**A**) Genome-wide gene (green) and TE (blue) density as well as a schematic representation of chromosomes (circle, centromere; dark gray, pericentromeric region; light gray, arm). (**B**) Genome-wide CG methylation divergence patterns among the 31st generation MA lines (MA1_1 and MA1_2) and the natural accessions (brown and gray, respectively). The red line indicates the theoretical prediction of divergence based on the estimated epimutation rates per annotation weighted by local annotation densities. Genome-wide divergence patterns in the MA lines are strongly correlated with the diversity patterns in the natural accessions (**C**), as well as with sequence diversity in the accessions (**D**). (**E**) The relationship between genetic and epigenetic variation as a function of time and different values of the epimutation rates α and β in a strictly selfing system without selection; x axis, time in generations; y axis, expected correlation between genotype and epigenotype of two perfectly linked loci (recombination fraction = 0). The estimated CG epimutation rate (blue line) is high enough to efficiently uncouple genetic from epigenetic variation over relatively short timescales (SI Appendix). (**F**) Equilibrium frequency (y axis) of epigenotypes $c^m c^m$ (solid) and $c^u c^u$ (dashed) in a strictly selfing system as a function of fitness (x axis) and different forward-backward epimutation rates (colored lines). The fitness of epigenotypes $c^u c^u$, $c^u c^m$ and $c^m c^m$ was defined by w, $0.5(1+w)$ and 1, respectively (SI Appendix).

Figure 2. (continued) … The estimated CG epimutation rate (blue line) is low enough to yield epimutation-selection equilibria close to those found for DNA sequence mutation rates, even under weak selection regimes (i.e., small fitness differentials between $c^u c^u$ and $c^m c^m$). This means that CG-type epialleles should be stable enough to effectively respond to long-term selection, provided they affect fitness.

[22] (SI Appendix, Table S8). We focused on a subset of 133 accessions that met our quality criteria and calculated CG-methylation diversity in a 1-Mb sliding window using the same protocol as with the MA lines (SI Appendix). Although the natural accessions were clearly more diverse (Fig. 2B), genome-wide diversity patterns were highly similar to those seen in the MA lines (weighted R^2 = 0.624, P < 0.0001, Fig. 2C and SI Appendix, Fig. S7), particularly in pericentromeric regions (R^2 = 0.899, P < 0.0001) and to a slightly lesser extent in chromosome arms (R^2 = 0.525, P < 0.0001). These observations are consistent with a recent report by Hagmann *et al.* [23]. Moreover, CG-methylation divergence among the MA lines was also moderately correlated with sequence diversity in the accessions, explaining over 25% of the genome-wide SNP distribution (weighted R^2 = 0.254, P < 0.0001, Fig. 2D and SI Appendix, Fig. S8).

It is unlikely that global patterns of CG-methylation diversity among natural accessions are the result of selection acting over broad genomic regions, because the same patterns are quickly established in isogenic MA lines in the course of only 31 generations under constant environmental conditions. Rather, our results suggest that these patterns reflect major structural properties of the *Arabidopsis thaliana* genome, which modulate the ratio of forward–backward epimutation rates, and thus determine the accumulation dynamics of neutral epimutations over time. It is therefore not surprising that the reorganization of genomes during macroevolution is necessarily accompanied by a repatterning of methylation divergence among lineages or species [24], insofar that such structural changes alter genome-wide annotation densities and their accompanying chromatin environment. However, structural changes of this type are less prevalent in the course of microevolution; hence, neutral epimutations are probably the single most important factor in shaping methylome diversity in populations over short to intermediate evolutionary timescales.

6.3 Discussion

6.3.1 CG epimutation rates are high enough to rapidly uncouple genetic and epigenetic variation over evolutionary timescales

Our analysis shows that CG epimutations are about five orders of magnitude more frequent than genetic mutations in *Arabidopsis thaliana* [~ 10^{-4} compared with ~ 10^{-9} [25]] and are subject to forward–backward dynamics that are rarely observed for genetic loci. Because of these properties, it is intuitively obvious that these epimutation dynamics will lead to an uncoupling of epigenetic from genetic variation over relatively short evolutionary timescales [26]. Simple deterministic models show that in a strictly selfing system without selection it would require only about 800 generations to reduce correlations between genotype and epigenotype from unity to below 0.5, and only about 2,700 generations to reduce it to below 0.1 (Fig. 2E and SI Appendix), and this breakdown is expected to be even faster in outcrossing systems. This rapid uncoupling may explain why variation in DNA methylation in *Arabidopsis thaliana* populations is only partly associated with *cis*- and *trans*-acting DNA sequence variants [22], and thus sheds new light on the molecular mechanisms that drive the coevolution of genomes and epigenomes. One situation in which genotype-epigenotype associations are expected to be more prevalent is when natural accessions have only newly diverged from a common ancestor, as it may be the case in recently founded local populations. This prediction can be tested using genome-wide association study-based *cis*- and *trans*-mapping analysis across different groups of accessions that vary along a gradient of genetic relatedness and (or) geographic locations. Recent whole-genome and methylome datasets of *Arabidopsis thaliana* local populations collected in North America [23] and Sweden [27] may be applicable for that purpose.

6.3.2 CG epimutation rates are low enough for new epialleles to sustain long-term selection responses

Although our results provide strong evidence that global patterns of CG-methylation diversity among *Arabidopsis thaliana* natural accessions are mainly influenced by the accumulation of selectively neutral epimutations, targeted selection of epialleles at specific loci may still be an important process. Particularly in chromosome arms, the MA divergence patterns were only moderately correlated with those of the natural accessions, suggesting the involvement of other factors such as direct selection on CG methylation states and (or) selection via DNA sequence variants that indirectly

regulate CG methylation in *cis* or *trans*. Indeed, the observation that methylation profiles of orthologous genes is often highly conserved across species [28] indicates that some epigenetic states are subject to strong evolutionary constraints. For epigenetic selection to be effective, epimutations need to be sufficiently stable [29], and a lack of stability has been cited as one reason why epigenetic inheritance has no potent role in evolution or in the heritability of complex traits [30]. Contrary to these conclusions, simple deterministic selection models show that newly arising epimutations are stable enough to respond effectively to long-term selection, even under weak selection regimes, yielding epimutation-selection equilibria that are close to those expected for DNA sequence mutation rates (Fig. 2F and SI Appendix).

6.3.3 Reference values for future population epigenetic studies

In light of our estimates of forward-backward epimutation rates, future work should examine the effect of selection in more complex population genetic models that account for finite population sizes, migration, and drift such as those proposed by Charlesworth and Jain [13]. Recently, Wang and Fan [14] devised a neutrality test based on single methylation polymorphism data using a modified version of Tajima's D. We caution that care needs to be taken when supplying epimutation rates to this or similar tests. Incorrect assumptions about the ratio of forward and backward rates can lead to widely misleading conclusions regarding the role of selection on CG methylation. If one assumes that forward and backward rates are equivalent, TE-associated CGs would most likely be detected as being under strong selection, and pericentromeric regions would seem to have undergone selective sweeps. However, if one considers that spontaneous methylation gain is about 30 times more likely than methylation loss (*see* Table 1), equilibrium levels of CG-methylation diversity in TEs would seem to be entirely consistent with neutrality. Hence, the context- or annotation-specific epimutation rates provided here should serve as useful reference values when inferring signatures of epigenetic selection in *Arabidopsis thaliana* and possibly in other plant species.

6.4 Materials and methods

Below we provide a brief description of the theoretical model and our estimation approach. For a more detailed explanation we refer the reader to SI Appendix.

6.4.1 Derivation of neutral epimutation model

Let c^u and c^m denote an unmethylated and a methylated cytosine, respectively, and $\alpha = Pr(c^u \rightarrow c^m)$ and $\beta = Pr(c^m \rightarrow c^u)$ be the probabilities that a cytosine gains or loses methylation during or before gamete formation, which can include gains or losses of DNA methylation in somatic tissues from which the gametic cells were derived. We arbitrarily call α the forward and β the backward epimutation rate per generation per haploid methylome. We modeled the epigenotype frequencies at the j^{th} cytosine using a Markov chain with three states: $c^u c^u$, $c^u c^m$, and $c^m c^m$. Taking into account Mendelian segregation of epialleles c^m and c^u together with rates α and β, we derived the epigenotype transition matrix $\mathbf{T}$ after one selfing generation:

	$\mathbf{c^u c^u}$	$\mathbf{c^m c^u}$	$\mathbf{c^m c^m}$
$\mathbf{c^u c^u}$	$(1 - \alpha)^2$	$2(1 - \alpha)\alpha$	α^2
$\mathbf{c^u c^m}$	$¼(\beta + 1 - \alpha)^2$	$½(\beta + 1 - \alpha)\ (\alpha + 1 - \beta)$	$¼(\alpha + 1 - \beta)^2$
$\mathbf{c^m c^m}$	β^2	$2(1 - \beta)\beta$	$(1 - \beta)^2$

This formulation does not account for higher-order epimutation events, because such events are expected to be rare for small epimutation rates. Following Markov chain theory, the epigenotype frequencies at cytosine j in the MA population after t generations of single seed descent, π_{tj} can be expressed as $\pi_{tj} = \pi_{0j} P \mathbf{V}^t P^{-1}$, where P is the eigenvector of matrix $\mathbf{T}$ and $\mathbf{V}$ is a diagonal matrix of the eigenvalues of matrix $\mathbf{T}$. Using Mathematica 10.0 (Wolfram Research, Inc.) we derived analytical solutions for the elements of π_{tj}, which are functions of t, α, β as well as the initial frequency vector π_{0j}. These analytical solutions have no easy form and are therefore omitted here for brevity. At equilibrium, the $\pi_{\infty j}$ represent the expected epigenotype frequencies at cytosine j among the MA lines after a (hypothetical) infinite number of selfing generations ($t = \infty$), and were obtained by calculating $lim_{t\rightarrow\infty}\pi_{tj}$:

$$\pi_{\infty j}(c^u c^u) = \frac{\beta\left((1-\beta)^2 - (1-\alpha)^2 - 1\right)}{(\alpha+\beta)\left((\alpha+\beta-1)^2 - 2\right)}$$

$$\pi_{\infty j}(c^u c^m) = \frac{4\alpha\beta(\alpha+\beta-2)}{(\alpha+\beta)\left((\alpha+\beta-1)^2 - 2\right)}$$

$$\pi_{\infty j}(c^m c^m) = \frac{\alpha\big((1-\alpha)^2 - (1-\beta)^2 - 1\big)}{(\alpha+\beta)\big((\alpha+\beta-1)^2 - 2\big)}$$

For any $0 < \alpha, \beta < 1$, these equilibrium solutions are independent of the initial epigenotype proportions π_{0j} in the common founder, and depend only on the rates α and β. The rate at which the epigenotype proportions converge to these equilibrium values depends on the relative magnitude of the forward and backward rates.

6.4.2 Modeling methylation divergence

To derive analytical formulas for methylation divergence, we score the methylation divergence between two independently selfed lines at every cytosine with the following distance matrix:

	$c^u c^u$	$c^m c^u$	$c^m c^m$
$c^u c^u$	0	½	1
$c^u c^m$	½	0	½
$c^m c^m$	1	½	0

Let t_1 and t_2 denote the number of generations between two individuals at generations G_m and G_n and their most recent common founder at generation G_f, respectively (i.e., $t_1 = G_m - G_f$, $t_2 = G_n - G_f$, Fig. 1A). Let $\pi_{t_i j}|c^m c^m$ be the vector of epigenotype frequencies at the j^{th} cytosine after t_i selfing generations from G_f, conditional on the fact that the most recent common founder epigenotype was $c^m c^m$: $\pi_{t_i j}|c^m c^m = (0,0,1) \cdot T^{t_i}$. Using this equation and the methylation divergence scoring table above, the divergence between these two lines at this locus can be calculated as

$$d_{t_1t_2j}|c^mc^m = \frac{1}{2}\sum_{k=1}^{4}(\pi_{t_1j}(P1_k)|c^mc^m \cdot \pi_{t_2j}(P2_k)|c^mc^m) + 1\sum_{k=1}^{2}(\pi_{t_1j}(Q1_k)|c^mc^m \cdot \pi_{t_2j}(Q2_k)|c^mc^m)$$

with $Q1 = \{c^uc^u, c^mc^m\}$, $Q2 = \{c^mc^m, c^uc^u\}$, $P1 = \{c^uc^u, c^uc^m, c^uc^m, c^mc^m\}$, and $P2 = \{c^uc^m, c^uc^u, c^mc^m, c^uc^m\}$. The simple multiplication of these frequencies follows from the fact that the selfing lines are conditionally independent. The divergence over all loci for which the most recent common founder at G_f was c^mc^m is

$$d_{G_f,t_1t_2}|c^mc^m = \sum_j d_{t_1t_2j}|c^mc^m = N_{G_f}^{mm} \cdot d_{t_1t_2j}|c^mc^m$$

where $N_{G_f}^{mm}$ are the number of methylated cytosines at G_f. The global (or total) DNA methylation divergence along the genome can be calculated as

$$D_{G_f,t_1t_2} = d_{G_f,t_1t_2}\left|c^mc^m + d_{G_f,t_1t_2}\right|c^uc^m + d_{G_f,t_1t_2}|c^uc^u$$

where $d_{G_f,t_1t_2}|c^uc^m$ and $d_{G_f,t_1t_2}|c^uc^u$ are derived using similar arguments as for $d_{G_f,t_1t_2}|c^mc^m$. We prefer to express the global methylation divergence as a proportion of all of the cytosines, in which case

$$D^*_{G_f,t_1t_2} = \frac{D_{G_f,t_1t_2}}{N}$$

Using the above derived equilibrium epigenotype frequencies, it can be shown that the equilibrium divergence is

$$D^*_\infty = \frac{2\alpha\beta\left(\left[(1-\beta)^2 - (1-\alpha)^2\right]^2 - 2[\alpha+\beta-1]^2 + 3\right)}{(\alpha+\beta)^2\left((\alpha+\beta-1)^2 - 2\right)^2}$$

6.4.3 Model fitting and parameter estimation

For each pedigree we had a number M of line comparisons and we denoted the observed methylation divergence between each of them as O_{G_f,t_1t_2i}, with $i = \{1,2,\ldots,M\}$, and G_f, t_1, and t_2 the times of and from their most recent common founder, respectively. We assumed that these observations were generated from the proposed epimutation model but contained some unknown measurement error. Hence, we had

$$O_{G_f,t_1t_2i} = c + D^*_{G_f,t_1t_2} + \epsilon_i$$

where c is the intercept, $D^*_{G_f,t_1t_2}$ is the theoretical global divergence measure introduced above, and ϵ is a random measurement error term. For the MA1_1 population the value of c was approximated using the methylation divergence between technical replicates. For the other three populations no technical replicates were available and c was estimated along with the other parameters (SI Appendix, Fig. S9). To obtain parameter estimates we minimized $r^2 = \sum_{\mathrm{i}}(O_{G_f,t_1t_2i} - D^*_{G_f,t_1t_2})^2$, which is a problem in multivariate nonlinear regression. This involves finding solutions to $\nabla r^2 = 0$, which can be obtained numerically. Extensive simulations showed that our estimation method performs well, even with relatively large measurement error (SI Appendix, Fig. S10).

Acknowledgments

We thank B. Charlesworth and J. Hadfield for their comments during a seminar at the University of Edinburgh. This work was supported by grants from the Netherlands Organization for Scientific Research (to R.C.J., R.W., A.v.d.G., F.J., and M.C.-T.), a University of Groningen Rosalind Franklin Fellowship (to M.C.-T.), National Institutes of Health Grant R00GM100000, and National Science Foundation Grant IOS-1339194 (to R.J.S.).

Supplementary material

Supplementary text, figures and tables can be found at
http://www.pnas.org/content/112/21/6676.long?tab=ds.

References

1. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**:204-220.
2. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**:245-249.
3. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334**:369-373.
4. Richards EJ (2006) Inherited epigenetic variation - Revisiting soft inheritance. *Nat Rev Genet* **7**:395-401.
5. Cortijo S, Wardenaar R, Colomé-Tatché M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury JM, Wincker P, Roudier F, Jansen RC, Colot V, Johannes F (2014) Mapping the epigenetic basis of complex traits. *Science* **343**:1145-1148.
6. Kalisz S, Purugganan MD (2004) Epialleles via DNA methylation: Consequences for plant evolution. *Trends Ecol Evol* **19**:309-314.
7. Weigel D, Colot V (2012) Epialleles in plant evolution. *Genome Biol* **13**:249.
8. Diez CM, Roessler K, Gaut BS (2014) Epigenetics and plant genome evolution. *Curr Opin Plant Biol* **18**:1-8.
9. Springer NM (2013) Epigenetics and crop improvement. *Trends Genet* **29**:241-247.
10. Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP (2014) Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. *Genome Res* **24**:1821-1829.
11. Becker C, Weigel D (2012) Epigenetic variation: Origin and transgenerational inheritance. *Curr Opin Plant Biol* **15**:562-567.
12. Hunter B, Hollister JD, Bomblies K (2012) Epigenetic inheritance: What news for evolution? *Curr Biol* **22**:R54-R56.
13. Charlesworth B, Jain K (2014) Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates. *Genetics* **198**:1587-1602.
14. Wang J, Fan C (2015) A neutrality test for detecting selection on DNA methylation using single methylation polymorphism frequency spectrum. *Genome Biol Evol* **7**:154-171.
15. Shaw RG, Byers DL, Darmo E (2000) Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. *Genetics* **155**:369-378.

16. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**:523-536.
17. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulfite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**:215-219.
18. Nuthikattu S, McCue AD, Panda K, Fultz D, DeFraia C, Thomas EN, Slotkin RK (2013) The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. *Plant Physiol* **162**:116-131.
19. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D (2013) The *Arabidopsis* nucleosome remodeler *DDM1* allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **153**:193-205.
20. Creasey KM, Zhai J, Borges F, Van Ex F, Regulski M, Meyers BC, Martienssen RA (2014) miRNAs trigger widespread epigenetically activated siRNAs from transposons in *Arabidopsis*. *Nature* **508**:411-415.
21. Yelagandula R, Stroud H, Holec S, Zhou K, Feng S, Zhong X, Muthurajan UM, Nie X, Kawashima T, Groth M, Luger K, Jacobsen SE, Berger F (2014) The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in *Arabidopsis*. *Cell* **158**:98-109.
22. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, Ecker JR (2013) Patterns of population epigenomic diversity. *Nature* **495**:193-198.
23. Hagmann J, Becker C, Müller J, Stegle O, Meyer RC, Wang G, Schneeberger K, Fitz J, Altmann T, Bergelson J, Borgwardt K, Weigel D (2015) Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. *PLoS Genet* **11**:e1004920.
24. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D (2014) Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS Genet* **10**:e1004785.
25. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**:92-94.
26. Johannes F, Colot V, Jansen RC (2008) Epigenome dynamics: A quantitative genetics perspective. *Nat Rev Genet* **9**:883-890.

27. Dubin MJ, Zhang P, Meng D, Remigereau MS, Osborne EJ, Paolo Casale F, Drewe P, Kahles A, Jean G, Vilhjálmsson B, Jagoda J, Irez S, Voronin V, Song Q, Long Q, Rätsch G, Stegle O, Clark RM, Nordborg M (2015) DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. *Elife* **4**:e05255.
28. Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci USA* **110**:1797-1802.
29. Furrow RE (2014) Epigenetic inheritance, epimutation, and the response to selection. *PLoS One* **9**:e101559.
30. Slatkin M (2009) Epigenetic inheritance and the missing heritability problem. *Genetics* **182**:845-850.

Summarizing discussion

Plant genomes make extensive use of DNA methylation for the silencing of transposable elements and the regulation of genes [1]. Before 2009, many reports had shown that the loss or gain of DNA methylation states can be stably inherited for many generations, and a number of studies even showed that spontaneous, single-locus DNA methylation variants (epialleles) can influence simple categorical traits such as flower shape (e.g. symmetrical versus asymmetrical flower; [2]). An emerging question at that time was whether the genome-wide segregation of multiple epialleles could provide a so far unexplored basis of variation for more "complex" traits, such as seed yield, drought resistance, biomass, flowering time, height, etc. This question was important because most of these latter traits are the primary targets of natural selection in the wild or artificial selection in agricultural breeding programs. However, there was no experimental system to answer these questions because commonly studied natural or experimental plant populations were rife with DNA sequence polymorphisms, which made it impossible to disentangle genetic from epigenetic effects.

To solve this problem, a panel of so-called epigenetic recombinant inbred lines (epiRILs) in the model plant *Arabidopsis thaliana* was established [3]. Individuals in this population have virtually identical DNA sequences but segregate hundreds of experimentally-induced differentially methylated regions (DMRs) across the genome. The epiRILs therefore provided a "clean" system to quantify the extent to which epigenetic variation can affect complex traits independently of DNA sequence polymorphisms. Indeed, extensive phenotypic analysis of the epiRILs showed that many ecologically and agriculturally important phenotypes are highly heritable in this population [3-5]. These phenotypic observations provided strong (but indirect) evidence that segregating DMRs are the cause for these heritability effects.

To provide a more direct demonstration of the causal link between DMRs and phenotypic variation in the epiRILs, we performed whole-genome methylation profiling for 123 epiRILs using methylated DNA immunoprecipitation in combination with tiling arrays (MeDIP-chip; [6, 7]). The epiRIL methylomes were reconstructed with the use of a three-state Hidden Markov Model (HMM). The choice for a HMM is based on two key properties of the MeDIP-chip data: (1) the signals of the probes are noisy proxies of an unobserved (or hidden) methylated, intermediate or unmethylated state and (2) the probes are spatially correlated and therefore

neighboring probes provide similar information. A HMM provides a powerful statistical framework for classifying individual probes given the overall structure of the data. Subsequent annotation analysis of the probe classifications shows that most of the gene probes are unmethylated and the majority of the transposable element probes are methylated, as expected.

--- Chapter 1 ---

In addition to MeDIP-chip we performed whole-genome bisulfite sequencing (WGBS-seq) for six of the 123 epiRILs for cross-validation purposes [7]. By conducting a side by side comparison between MeDIP-chip and WGBS-seq we show that MeDIP-chip performs reasonable well in detecting DNA methylation at the probe level, yielding a genome-wide combined false-positive and false-negative rate of about 0.21 [8]. Results however indicate that the detection can be susceptible to strong signal distortions resulting from a combination of dye bias and the CG content of effectively unmethylated genomic regions. Nonetheless, we show that these issues can be easily bypassed by taking appropriate data preparation steps and applying suitable analysis tools. Altogether we conclude that MeDIP-chip is a reasonable alternative to WGBS-seq for the detection of DMRs.

--- Chapter 2 ---

A direct view of the full methylomes of the epiRILs revealed that many DMRs segregate in a strictly Mendelian fashion for at least eight generations while others display highly dynamic (non-Mendelian) changes. Using these stable DMRs as physical markers, we were able to construct a recombination map in this isogenic population [7]. This was the first time that anyone had access to the recombination landscape of an experimental population that was variable at the DNA methylation level but essentially free of segregating DNA sequence polymorphisms. With this system we were in a unique position to answer novel questions regarding the role of genetic and epigenetic variation in shaping the recombination landscape in *Arabidopsis*. Prior to this work, both DNA methylation and sequence polymorphisms were believed to be major barriers of crossing-over events, particularly in pericentromeric regions of chromosomes. Contrary to this expectation, we found that suppression of recombination was strongly maintained in core pericentromeres in the epiRILs, and even further reinforced at pericentromeric boundaries. As a trade-off, crossing-over frequencies in chromosome arms were increased, so that the total number of crossing-over events was largely preserved globally. From these results, we concluded that neither dense DNA methylation nor DNA sequence polymorphisms are major determinants of the suppression of recombination in

pericentromeric regions, and that other, yet unknown, factors are involved. These results were consistent with other reports that were published nearly concurrently, but using widely different experimental approaches.

--- Chapter 3 ---

We employed the above-mentioned recombination map in conjunction with classical linkage mapping to identify segregating DMRs that could explain the heritable phenotypic effects previously observed in the epiRILs [9]. Using this approach, we detected several chromosomal regions that harbor causal DMRs, which together accounted for up to 90% of the heritability in flowering time and root length. This was the first demonstration, in any organisms, that DMRs can be stably inherited independently of DNA sequence changes and function as so-called epigenetic quantitative trait loci (QTL^{epi}). Phenotypically, the detected QTL^{epi} have all the necessary properties to become targets of natural or artificial selection. This possibility has implications for evolutionary biology and applied agricultural genetics, and has opened entirely new research lines in the emerging field of population epigenetics and epigenomics.

--- Chapter 4 ---

One important phenomenon that has been extensively exploited in agriculture is called heterosis (also called hybrid vigor). Heterosis describes an F1 phenotype with improved or increased performance compared to its parent varieties. Despite its importance for crop improvement the knowledge of the mechanisms behind this phenomena remain incomplete. Interestingly, there is growing evidence that epigenetic divergence plays an important role [10-12]. In order to test whether divergence in DNA methylation can contribute to heterosis we constructed epigenetic F1 hybrids (epiHybrids) by crossing a Col-0 wild-type as maternal parent with 19 different epiRILs (Chapter 1 and 3) as paternal parent [13]. Both parental epiRIL and wild-type lines were nearly isogenic making it an ideal system to explore the contribution of DNA methylation divergence to heterosis. We focused on seven different traits (leaf area, growth rate, flowering time, main stem branching, rosette branching, final plant height and seed yield) and observed a remarkable wide range of heterotic effects. Sixteen of the 19 epiHybrids exhibited significant mid-parent heterosis and among these nine exhibited significant low- or high-parent heterosis for at least one of the seven traits. Up to 51 % of the total variation in mid-parent divergence could be attributed to (epi)genomic differences between the Col-0 wild-type and epiRILs used for the crosses. Furthermore with the use of classical linkage analysis we detected several QTL for leaf area, flowering time and height that could

potentially explain the observed variation in mid-parent heterosis. Several of the candidate DMRs detected within the QTL regions are close to potentially interesting candidate genes. All together these results could potentially have implications for future crop breeding.

--- Chapter 5 ---

The epiRIL study, as well as several other studies have shown examples of segregating epimutations (in experimental populations or natural populations) that in some cases contribute to heritable phenotypic variation. These observations have led to much speculation about the role of DNA methylation in plant evolution, and its potential in breeding programs. Precise estimates of the rates at which methylation is spontaneously gained (forward epimutation) or lost (backward epimutation) at individual cytosines are necessary to be able to quantify the long-term dynamics of epigenetic variation under laboratory or natural conditions, and to understand the molecular mechanisms that drive methylome evolution. In order to obtain these rates we combined theoretical modelling with high-resolution methylome analysis of multiple independent mutation accumulation (MA) lines [14]. Results show that the epimutation rates of CG cytosines are about five orders of magnitude higher compared to genetic mutations (10^{-4} compared to 10^{-9}). Moreover, the rate at which methylation is lost ($6.30 \cdot 10^{-4}$) is about three times higher compared to the rate at which methylation is gained ($2.56 \cdot 10^{-4}$). Annotation-specific epimutation rates were in the same order but subtle differences in the rates had a substantial impact on the divergence of CG methylation across annotations. Furthermore results indicate that these context specific rates have shaped the global patterns of methylation divergence in *Arabidopsis* natural populations rather than selection taking place over broad genomic regions, because the patterns in the MA lines, which show a high correlation with the patterns of the natural accessions, were established in only 31 generations. Finally, theoretical arguments indicate that the epimutation rates are strong enough to rapidly uncouple genetic from epigenetic variation, but low enough for new epimutations to sustain long-term selection responses. The obtained results provide new insights into methylome evolution and its population-based consequences.

--- Chapter 6 ---

References

1. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**:204-220.
2. Cubas P, Vincent C, Coen E (1999) An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401**:157-161.
3. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, Bouchez D, Dillmann C, Guerche P, Hospital F, Colot V (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5**:e1000530.
4. Latzel V, Zhang Y, Karlsson Moritz K, Fischer M, Bossdorf O (2012) Epigenetic variation in plant responses to defense hormones. *Ann Bot* **110**:1423-1428.
5. Roux F, Colomé-Tatché M, Edelist C, Wardenaar R, Guerche P, Hospital F, Colot V, Jansen RC, Johannes F (2011) Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. *Genetics* **188**:1015-1017.
6. Cortijo S, Wardenaar R, Colomé-Tatché M, Johannes F, Colot V (2014) Genome-wide analysis of DNA methylation in *Arabidopsis* using MeDIP-chip. *Methods Mol Biol* **1112**:125-149.
7. Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E, Labadie K, Wincker P, Jacobsen SE, Jansen RC, Colot V, Johannes F (2012) Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci USA* **109**:16240-16245.
8. Wardenaar R, Liu H, Colot V, Colomé-Tatché M, Johannes F (2013) Evaluation of MeDIP-chip in the context of whole-genome bisulfite sequencing (WGBS-seq) in *Arabidopsis*. *Methods Mol Biol* **1067**:203-224.
9. Cortijo S, Wardenaar R, Colomé-Tatché M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury JM, Wincker P, Roudier F, Jansen RC, Colot V, Johannes F (2014) Mapping the epigenetic basis of complex traits. *Science* **343**:1145-1148.
10. Groszmann M, Greaves IK, Fujimoto R, Peacock WJ, Dennis ES (2013) The role of epigenetics in hybrid vigour. *Trends Genet* **29**:684-690.
11. Springer NM (2013) Epigenetics and crop improvement. *Trends Genet* **29**:241-247.

12. Dapp M, Reinders J, Bédiée A, Balsera C, Bucher E, Theiler G, Granier C, Paszkowski J (2015) Heterosis and inbreeding depression of epigenetic *Arabidopsis* hybrids. *Nat Plants* **1**:15092.
13. Lauss K, Wardenaar R, van Hulten M, Guryev V, Keurentjes JJ, Stam M, Johannes F (2016) Epigenetic divergence is sufficient to trigger heterosis in *Arabidopsis thaliana*. [preprint]
14. van der Graaf A, Wardenaar R, Neumann DA, Taudt A, Shaw RG, Jansen RC, Schmitz RJ, Colomé-Tatché M, Johannes F (2015) Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proc Natl Acad Sci USA* **112**:6676-6681.

Nederlandse samenvatting (Dutch summary)

Het is al geruime tijd bekend dat genetische variatie (variatie in de genetische code) **in** het DNA zit, overgeërfd kan worden en kan bijdragen aan de variatie van eigenschappen binnen de populatie (bijv. de lengte van individuen). Wat minder bekend is en wat een fundamentele vraag is binnen de wetenschap, is of de variatie die **op** het DNA ligt (welke een rol heeft in de organisatie en regulatie van het DNA) ook overgeërfd kan worden en kan bijdragen aan de variatie in eigenschappen die je ziet in een populatie los van de genetische code zelf. Men spreekt in dit geval van epigenetische variatie (epi: Grieks voor 'boven', 'op' of 'over') en het vakgebied van de epigenetica. Epigenetische variatie wordt onderverdeeld in drie categoriën; (1) een chemische verandering van het DNA (methylering van de base cytosine is het meest bestudeerd), (2) een chemische verandering van de eiwitten (histonen) die samen met het DNA chromatine vormen (bijv. acetylering of andere vormen van chemische modificaties) en (3) de aanwezigheid van kleine interferentie-RNAs. In dit proefschrift richten we ons op DNA methylering (toevoeging van een methylgroep aan de base cytosine) en bestuderen we deze vorm van epigenetische variatie in de model plant *Arabidopsis thaliana* (zandraket).

Plantengenomen maken extensief gebruik van DNA methylering om de activiteit van transposons (transcriptie en transpositie) te remmen en om de activiteit van genen (transcriptie) te reguleren. Voor 2009 hebben verschillende raporten aangetoond dat de mate van methylering (wel of geen methylatie) stabiel kan worden overgeërfd, en een aantal studies heeft zelfs aangetoond dat spontane methyleringsvariatie op één locus (een epi-allel) invloed kan hebben op eenvoudige eigenschappen zoals de vorm van een bloem (bijv. symmetrische ten opzichte van asymmetrische bloemvorm). Een opkomende vraag in die tijd was of de overerving van deze epi-allelen over het hele genoom kan bijdragen aan een tot nu toe nog niet ontdekte basis van "complexere" eigenschappen, zoals zaadopbrengst, droogteresistentie, biomassa, bloeitijd (Engels: flowering time), hoogte etc. Deze vraag was belangrijk omdat deze laatste eigenschappen doelwitten (targets) zijn voor natuurlijke selectie in het wild of kunstmatige selectie in agrarische teeltprogramma's. Er was echter nog geen experimenteel systeem om deze vragen te beantwoorden omdat de gebruikelijke natuurlijke of experimentele plantpopulaties naast variatie in DNA methylatie ook DNA sequentie verschillen bevatten. Hierdoor is het moeilijk om genetische van epigenetische effecten te onderscheiden (bijdrage van beide aan de eigenschappen van de plant). Als

oplossing hebben we een populatie van zogenoemde epigenetische recombinante inteelt lijnen (epiRILs) geïntroduceerd in de model plant *Arabidopsis thaliana* (*A. thaliana*). De individuele planten hebben praktisch identieke DNA sequenties maar segregeren honderden experimenteel-geïnduceerde differentieel gemethyleerde regio's (Engels: Differentially Methylated Regions; DMRs) verspreid over het hele genoom. De epiRILs bieden daarom een "zuiver" systeem om te kwantificeren in welke mate epigenetische variatie kan bijdragen aan variatie in complexe eigenschappen onafhankelijk van (mogelijke) verschillen in de DNA sequentie. Inderdaad, uitgebreid fenotypisch onderzoek van de epiRILs heeft aangetoond dat verschillende ecologisch and agrarisch belangrijke eigenschappen in hoge mate erfelijk zijn in deze populatie. Deze fenotypische observaties maken het aannemelijk dat segregerende DMRs deze erfelijke effecten veroorzaken.

In dit proefschrift gaan we verder met deze zoektocht naar de erfelijke basis van DNA methylatie voor de variatie van complexe eigenschappen. We hebben de methylatiepatronen bepaald voor het gehele genoom voor meer dan 100 epiRILs (beschreven in **hoofdstuk 1** en **2**). Dit was gedaan met behulp van immunoprecipitatie van gemethyleerde DNA fragmenten gevolgd door hybridisatie op micro-chips en bisulfiet behandeling van gefragmenteerd DNA gevolgd door next-generation sequencing (NGS) technologiën. Na het bepalen van de methylatiepatronen waren we in staat om een recombinatiekaart te maken voor de epiRILs in deze populatie (beschreven in **hoofdstuk 3**). Deze kaart was gemaakt met behulp van DMRs die stabiele segregatie (volgens Mendel) vertonen zoals verwacht op basis van de kruising die was toegepast. Dit was de eerste keer in de geschiedenis dat men toegang had tot de recombinatiepatronen in een populatie die variabel is op DNA methylatieniveau maar, in essentie, vrij is van segregerende DNA sequentieverschillen. Deze recombinatiekaart hebben we vervolgens gebruikt voor klassieke linkage analyse voor het vinden van DMRs die de geobserveerde fenotypische variatie zouden kunnen verklaren (beschreven in **hoofdstuk 4**). Met behulp van deze aanpak waren we in staat om diverse chromosomale regio's te detecteren die DMRs bevatten, die samen tot 90 % van de erfelijkheid verklaren voor de gemeten bloeitijd en wortellengte in deze populatie. Dit was de eerste demonstratie, in enig organisme, dat DMRs stabiel overgeërfd kunnen worden onafhankelijk van DNA sequentieverschillen en functioneren als zogenoemde epigenetische quantitative trait loci (QTL^{epi}). We realiseerden ons dat de epiRILs ook gebruikt konden worden voor het bestuderen van andere fenomenen zoals heterosis (ook wel hybrid vigor genoemd), een fenomeen waarbij de fenotypische eigenschappen van F1 hybride nakomelingen betere eigenschappen (bijvoorbeeld

meer vruchten) vertonen dan de ouders. Dit fenomeen is veelal benut in agrarische teeltprogramma's voor verscheidene eeuwen als een manier om gewasopbrengst te maximaliseren en rassen te verbeteren (bijvoorbeeld melkkoeien). Voor dit doeleinde hebben we verschillende hybridepopulaties gecreëerd door epiRILs te kruisen met wild-type *A. thaliana* planten (beschreven in **hoofdstuk 5**). Verschillende eigenschappen werden gemeten zoals bladoppervlak en de hoogte van de planten. We hebben met deze experimentele setting bewijs verkregen die aantonen dat erfelijke verschillen in cytosine methylatie tussen twee ouderlijke *Arabidopsis* lijnen genoeg zijn om heterosis teweeg te brengen in hun F1 nakomelingen zelfs in de nagenoege afwezigheid van DNA sequentie verschillen.

Hoewel we met onze aanpak met de epiRILs veel experimentele controle hebben, geeft het geen goed beeld van hoe de DNA methylatieverschillen in natuurlijke settings worden verkregen. In *A. Thaliana* heeft men aangetoond dat epimutaties (verandering in de methylatie toestand van CG dinucleotiden) spontaan ontstaan en in een snelle en niet-lineaire manier accumuleren, zeer waarschijnlijk als een resultaat van snelle toe- en afname dynamiek (Engels: forward-backward dynamics; verwerven van methylatie en verliezen van methylatie). Als een uitbreiding op ons werk met de epiRILs hebben we openbare en nieuw-gegenereerde bisulfite sequencing data gebruikt om robuuste schattingen te krijgen van de snelheid waarmee spontaan methylatie toeneemt (forward epimutatie) of afneemt (backward epimutatie) op het niveau van individuele cytosines (beschreven in **hoofdstuk 6)**. De resultaten laten zien dat de snelheden afhankelijk zijn van de genoomcontext (gen regio's, transposon regio's, etc.) en dat deze contextverschillen het epimutatielandschap langs het gehele genoom hebben vormgegeven in *A. thaliana*. We hebben ook gedemonstreerd dat de kennis van deze snelheden gebruikt kan worden voor het voorspellen van patronen van methyloomdivergentie over tijdschalen die van agrarisch en evolutionair belang zijn.

Alles bij elkaar genomen heeft het werk dat gepresenteerd is in dit proefschrift substantiële bijdrage geleverd aan het snelgroeiende vakgebied van de plant epigenetica.

List of abbreviations and acronyms

AT	Annotation transition (zone)
BLAST	Basic local alignment search tool
bp	Base pair(s)
BS	Bisulfite
ChIP	Chromatin immunoprecipitation
ChIP-chip	ChIP followed by hybridization on a tiling array
ChIP-seq	ChIP followed by next-generation sequencing
cM	CentiMorgan
chr	Chromosome
CO	Crossing over
DAS	Days after sowing
DDM1 (*ddm1*)	Decrease in DNA methylation 1
DMR	Differentially methylated region
DS	Dye-swap
epiHybrid	Epigenetic (F1) hybrid
epiRIL	Epigenetic recombinant inbred line
FN	False-negative
FP	False-positive
FT (FT1, FT2)	Flowering time
GR	Growth rate
GWFP	Genome-wide false-positive (rate)
G/R	IP DNA labelled with Cy3 (green) and input DNA with Cy5 (red)
HT	Final plant height
IP	Immunoprecipitated DNA
LA	Leaf area
LD	Linkage disequilibrium
LOD	Logarithm of the odds ratio for linkage
LPH	Low-parent heterosis
HMM	Hidden Markov model
HPH	High-parent heterosis
MA	Mutation accumulation
Mb	Mega base pairs = 1,000,000 bp
MeDIP	Methyl (or methylated) DNA immunoprecipitation
MeDIP-chip	MeDIP followed by hybridization on a tiling array
MeDIP-seq	MeDIP followed by next-generation sequencing

MP	Mid-parent
MPV	Mid-parent value
MSB	Main stem branching
NGS	Next-generation sequencing
nt	Nucleotide
PCR	Polymerase chain reaction
QTL^{epi}	Epigenetic quantitative trait locus (or loci)
QTL	Quantitative trait locus (or loci)
RB	Rosette branching
RGB	Additive color model in which red, green and blue are added in various ways to reproduce a broad array of colors.
R/G	IP DNA labelled with Cy5 (red) and input DNA with Cy3 (green)
RL (RL1, RL2)	Primary root length
SE	Standard error
SEE	SE of the estimate
SEM	SE of the mean
siRNA	Small interfering RNA
SMP	Single methylation polymorphism
SY	Seed yield
TAIR	The *Arabidopsis* Information Resource; a community resource and online database of genetic and molecular biology data for the model plant *Arabidopsis thaliana*
TCM	Trans chromosomal methylation
TCdM	Trans chromosomal demethylation
TE	Transposable element or Tris - EDTA (TE) buffer
TEASV	TE-associated structural variant
URL	Uniform resource locator; a reference (an address) to a resource on the internet
WGBS-seq	Whole-genome bisulfite sequencing
WT (wt)	Wild-type

Acknowledgments

After being part of the Groningen Bioinformatics Centre (GBIC) for about eight years, of which the last three years as a PhD student, it feels good to finish my time as a PhD student! One of the things I never envisioned during my time at GBIC is that I would get the opportunity and would decide to do a PhD. There were many people involved along the road to this thesis. In special I would like to express my gratitude to the persons mentioned below.

I am greatly indebted to my promotor **Prof. Dr. Ritsert C. Jansen** for the opportunity to do a PhD at his group. I would like to thank you for all the conversations we had over the years. Whether it was work related or private it was always pleasant to talk. Thanks for clearing my view on things. Thank you for your constructive comments, your inspiration and support during the years and also with finalizing this thesis!

My special gratitude goes to **Dr. Frank Johannes**, my copromotor and daily supervisor. Starting from my internship in 2008 till now the end of my PhD; It has been a long journey together. Thanks for coaching me and being patient with me for all those years! I enjoyed traveling together (to meetings, conferences, etc.). Thanks for introducing me to new people and for getting me involved in the different projects. Also, I would like to thank you for your feedback, your help with writing and presentations and the meaningful conversations. Thanks for pushing me forward whenever I needed it!

I also in special would like to thank **Klazien Offens**, our secretary. Thanks for helping us (and me) with all the paper work, organization of social events and things that needed to be arranged for a PhD defense and thesis. Even after changing to a different department you were offering to help us which I, and I think we all, really appreciate. You were, and still are, a valuable asset of the group!

I also would like to thank the members of the reading committee; **Prof. Dr. L. H. Franke**, **Prof. Dr. L. W. Beukeboom** and **Prof. Dr. O. Bossdorf** for the critical assessment of this thesis.

The two colleagues with whom I shared my office room, **Maria Colomé-Tatché** and **Pariya Behrouzi**. **Maria**, thanks for the time that we have been working together and thanks for sharing the office with me! I always experienced it as a very good and

productive time. Also thank you for your critical view on things and thanks for the scientific and non-scientific conversations (life in general). **Pariya**, also thanks for sharing the office with me! I enjoyed our conversations during lunch or in the office when it was about personal things or work. It was also nice to see how you are starting to integrate into the Dutch culture (Dutch classes; swimming; etc.)!

To my colleague in epigenetics, **Lionel Morgado**: Thanks for sharing your passion for small RNAs and thanks for your critical view on things. It was sometimes hard to leave the office and to catch my bus when I was stuck in a conversation with you! But no regrets; It was about some good and interesting science! I enjoyed our time together at GBIC which now brought us to Germany.

To my housemate and colleague, **Aaron Taudt**: Thanks for giving me the opportunity to stay at your apartment! The timing was really convenient. I also enjoyed our time together outside the office hours (with David, Amaryllis, Lionel, and other people). I think I still need to improve my board game skills. You are hard to beat!

I also wish to express my thanks to my colleague **Konrad Zych**. Thanks for your help with software (viruses, PDF conversion, etc.) and hardware (I am still very happy with my 6T hard drive!)! I enjoyed listening to your talks. Thanks for broadening my view on potatoes!

I also would like to say thanks to my former colleague, **Adriaan van der Graaf**. Thanks for all the scientific and non-work related conversations. It was also nice to speak Dutch once in a while! Also thanks for the conversations we had outside the office. I am not really a drinker, but thanks for showing me some of the best bars in Groningen!

Haiyin Liu, thanks for your work on the MeDIP-chip and WGBS-seq data (Evaluation of MeDIP-chip in context WGBS-seq) which resulted in a chapter in Methods in Molecular Biology!

The latest members of the JohannesLab, **David Roquis** and **Amaryllis Vidalis**. It was great to get to know both of you. I enjoyed our time inside and outside the office (board games, drinks). I am not really a drinker, but it is hard to refuse a good cocktail! Also thanks for using the bisulfite pipeline. It still needs some fine-tuning and it is not really user-friendly yet… but we will get there…! Thanks for using, testing

and thanks for your feedback, ideas and all the conversations that we have had together!

I would also like to thank all the people that were part of GBIC over the years and left the group or moved to the UMCG before the finishing of my PhD, like **Danny Arends**, **Joeri van der Velde**, **Morris Swertz**, **Elena Merlo**, **Mohammed Shojaei Arani**, **Minh Anh Nguyen**, **Marnix Medema**, **Richard Scheltema**, **Yang Li**, **Bruno Tesson**, **Anna Kolesnichenko**, and many others. It was great meeting you and you were a great inspiration to me.

Peter Terpstra, thanks for helping us with the teaching of the Bioinformatics course!

I would also like to thank **Victor Guryev** from the European Research Institute for the Biology of Ageing (ERIBA). I would like to thank you for our collaboration that led to the recent Nature Communications paper!

I would also like to express my gratitude to **Vincent Colot** and **Sandra Cortijo**, our collaboration partners from IBENS (CNRS, Paris). Both I would like to thank for the time that we worked together on projects related to *Arabidopsis* epiRILs. It was great to meet you and to have scientific conversations. It was really a great privilege. **Vincent**, I would like to thank you again for the opportunity to visit your group in 2011. It was a great experience and I have learned a lot.

I would also in special like to thank **Maike Stam** and **Kathrin Lauss**, our collaboration partners from the University of Amsterdam. It was, and still is, a privilege to work with you. Thanks for all the scientific discussions related to epiHybrids. Thanks for introducing me to the phenomenon of heterosis! **Kathrin**, also thanks for showing me your lab. It is nice to see how we all contribute in different ways!

Ik wil ook graag mijn **ouders** en **familie** bedanken voor de steun voor de afgelopen jaren. Vooral de laatste loodjes waren zwaar. Bedankt voor al jullie steun, zorg en liefde! Bedankt dat ik bij jullie terecht kon!

Als laatste wil ik **God** bedanken voor inspiratie, kracht en alle hulp door alle jaren heen!

René Wardenaar
Freising, Germany, 6th of November 2016

About the author

Curriculum vitae

René Wardenaar was born on the 5th of August 1984 in Leek, a town located in the province of Groningen, The Netherlands. In 2003 he started to study Chemistry at the University of Groningen (Rijksuniversiteit Groningen) where he completed the first year. During this time his interest in Biology and a more "dry" way of doing research (working with computers) was growing. This resulted in a switch to the study of Bioinformatics at the Hanze University of Applied Sciences (Hanzehogeschool Groningen) in 2006. The study of Bioinformatics was a good combination of Biology, Biochemistry, Genetics and other subjects related to Biological Sciences in combination with "dry" research work (programming, statistical analysis, etc.).

In the final year of his study of Bioinformatics, he was offered an internship and a graduation project at the Groningen Bioinformatics Centre (GBIC). GBIC was led by Prof. Dr. Ritsert C. Jansen and was located in Haren at that time. During the internship he participated in the development of several statistical models for the detection of chromatin differences (e.g. DNA methylation) between different cell types (e.g. normal versus cancerous cells). This project resulted in the author's first paper. The research topic of his graduation project was about the inheritance of DNA methylation patterns in hematopoietic stem cells in mice. His work on hematopoietic stem cells was awarded with a shared ***second place for best undergraduate thesis*** in 2009 by the Royal Netherlands Chemical Society (KNCV; Koninklijke Nederlandse Chemische Vereniging). Both internship and graduation project were under the supervision of Dr. Frank Johannes who is currently an Assistant Professor at the Technical University of Munich.

After finishing his bachelor in Bioinformatics he continued working at GBIC as a research assistant (under supervision of Dr. Frank Johannes). This resulted in several publications on topics that are related to DNA methylation inheritance in mammals as well as in the model plant *Arabidopsis thaliana*. The majority of the projects were

based on work with epigenetic recombinant inbred lines (epiRILs) in *Arabidopsis thaliana*. EpiRILs are RILs that have only a few sequence differences (near isogenic) but highly divergent methylomes; many differences in DNA methylation. This population of plants was created by a team that was led by Dr. Vincent Colot (Institut de Biologie de l'École Normale Supérieure, CNRS; Paris). In 2011 the author was invited to join the Colot Lab in Paris for two periods of about two months each. He gratefully accepted both invitations. The time in Paris resulted in the construction of a pipeline for the analysis of bisulfite sequencing data which was subsequently used for the reconstruction of several epiRIL methylomes. Altogether, the work with the epiRIL population resulted in a ***Science paper*** in 2014. In this paper it was demonstrated that differentially methylated regions (DMRs) can contribute to phenotypic variation independently from sequence variation and act as bona fide epigenetic quantitative trait loci (QTL^{epi}) in this epiRIL population. The author was also involved in a project focusing on heterosis in epigenetic hybrids which were constructed with the use of 19 epiRILs. This project was in collaboration with the research group led by Dr. Maike Stam (University of Amsterdam). In 2016 the author spent about half a year at the group of Dr. Frank Johannes who established a research group in Population Epigenetics and Epigenomics at the Technical University of Munich, Germany. In this period the author was working on the improvement of an already existing pipeline for the analysis of bisulfite sequencing data and the above mentioned heterosis project (collaboration with the Stam Lab).

During his time at GBIC the author gratefully accepted a position as a PhD student (in 2013; under the supervision of Prof. Dr. Ritsert C. Jansen and Dr. Frank Johannes). The time at GBIC resulted in this PhD thesis.

From December 2016 on, the author will be working as a postdoctoral researcher at the group of Dr. Frank Johannes.

Journal publications

* Shared first author

- Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, **Wardenaar R**, Renkens I, Coe BP, Deelen P, de Ligt J, Lameijer EW, van Dijk F, Hormozdiari F, Genome of the Netherlands Consortium, Uitterlinden AG, van Duijn CM, Eichler EE, de Bakker PI, Swertz MA, Wijmenga C, van Ommen GB, Slagboom PE, Boomsma DI, Schönhuth A, Ye K, Guryev V (2016) A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**:12989.

- van der Graaf A*, **Wardenaar R***, Neumann DA, Taudt A, Shaw RG, Jansen RC, Schmitz RJ, Colomé-Tatché M, Johannes F (2015) Rate, spectrum and evolutionary dynamics of spontaneous epimutations. *Proc Natl Acad Sci USA* **112**(21), 6676-6681.

- Kooke R, Johannes F, **Wardenaar R**, Becker F, Etcheverry M, Colot V, Vreugdenhil D, Keurentjes JJ (2015) Epigenetic basis of morphological variation and phenotypic plasticity in *Arabidopsis thaliana*. *Plant cell* **27**(2), 337-348.

- Cortijo S*, **Wardenaar R***, Colomé-Tatché M*, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury JM, Wincker P, Roudier F, Jansen RC, Colot V, Johannes F (2014) Mapping the epigenetic basis of complex traits. *Science* **343**(6175), 1145-1148.

- Cortijo S*, **Wardenaar R***, Colomé-Tatché M, Johannes F, Colot V (2014) Genome-wide analysis of DNA methylation in *Arabidopsis* using MeDIP-chip. *Methods Mol Biol* **1112**, 125-149.

- **Wardenaar R**, Liu H, Colot V, Colomé-Tatché M, Johannes F (2013) Evaluation of MeDIP-chip in the context of whole-genome bisulfite sequencing (WGBS-seq) in *Arabidopsis*. *Methods Mol Biol* **1067**, 203-224.

- Lendvai A, Johannes F, Grimm C, Eijsink JJ, **Wardenaar R**, Volders HH, Klip HG, Hollema H, Jansen RC, Schuuring E, Wisman GB, van der Zee AG (2012) Genome-wide methylation profiling identifies hypermethylated biomarkers in high-grade cervical intraepithelial neoplasia. *Epigenetics* **7**(11), 1268-1278.

- Colomé-Tatché M, Cortijo S, **Wardenaar R**, Morgado L, Lahouze B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E, Labadie K, Wincker P, Jacobsen SE, Jansen RC, Colot V, Johannes F (2012) Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci USA* **109**(40), 16240-16245.

- Roux F, Colomé-Tatché M, Edelist C, **Wardenaar R**, Guerche P, Hospital F, Colot V, Jansen RC, Johannes F (2011) Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. *Genetics* **188**(4), 1015-1017.

- Kooistra SM, van den Boom V, Thummer RP, Johannes F, **Wardenaar R**, Tesson BM, Veenhoff LM, Fusetti F, O'Neill LP, Turner BM, de Haan G, Eggen BJ (2010) Undifferentiated embryonic cell transcription factor 1 regulates ESC chromatin organization and gene expression. *Stem Cells* **28**(10), 1703-1714.

- Johannes F*, **Wardenaar R***, Colomé-Tatché M, Mousson F, de Graaf P, Mokry M, Guryev V, Timmers HT, Cuppen E, Jansen RC (2010) Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* **26**(8), 1000-1006.

Under review / Preprint / In preparation

- Vidalis A, Živković D, **Wardenaar R**, Roquis D, Tellier A, Johannes F. Methylome evolution in plants. [*Under review*]

- Lauss K, **Wardenaar R**, van Hulten M, Guryev V, Keurentjes JJ, Stam M, Johannes F. Epigenetic divergence is sufficient to trigger heterosis in *Arabidopsis thaliana*. [*Preprint*]

- **Wardenaar R**, Johannes F. Stable DMRs at *ddm1-2* epimutable sites are largely independent of standing genetic variation in *Arabidopsis thaliana* natural populations. [*In preparation*]

Presentations

- *The 9th Annual NBIC Conference*, Lunteren, The Netherlands, April 2014:
  Mapping the epigenetic basis of complex traits.

- *NBIC BioRange Project Meeting*, Ede, The Netherlands, October 2012:
  Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation.

Posters

- *The 8th Annual NBIC Conference*, Lunteren, The Netherlands, April 2013:
  Wardenaar R, Liu H, Colot V, Colomé-Tatché M, Johannes F. Evaluation of MeDIP-chip in the context of Whole Genome Bisulfite Sequencing (WGBS-seq) in *Arabidopsis*.

- *The 20th Annual GBB Symposium*, Groningen, The Netherlands, September 2012:
  Colomé-Tatché M, Cortijo S, **Wardenaar R**, Lahouze B, Etcheverry M, Martin A , Feng S, Duvernois-Berthet E, Labadie K, Wincker P, Jacobsen SE, Jansen RC, Colot V, Johannes F. Recombination patterns are resistant to extensive loss of sequence variation and DNA methylation in *Arabidopsis*.

- *The 7th Annual NBIC Conference*, Lunteren, The Netherlands, April 2012:
  Colomé-Tatché M, Cortijo S, **Wardenaar R**, Lahouze B, Etcheverry M, Martin A , Feng S, Duvernois-Berthet E, Labadie K, Wincker P, Jacobsen SE, Jansen RC, Colot V, Johannes, F. Recombination patterns are resistant to extensive loss of sequence variation and DNA methylation in *Arabidopsis*.

- *NBIC BioRange Project Meeting*, Ede, The Netherlands, October 2011:
  Cortijo S, Colomé-Tatché M, **Wardenaar R**, Jacobsen SE, Wincker P, Jansen RC, Johannes F, Colot V. Genome-wide analysis of inheritance patterns of DNA methylation reveals widespread epiallele stability in *Arabidopsis*.

- *The 18th Annual GBB Symposium*, Groningen, The Netherlands, September 2010:
  Johannes F, **Wardenaar R**, Colomé-Tatché M, Mousson F, de Graaf P, Mokry M, Guryev V, Timmers HT, Cuppen E, Jansen RC. Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq.

- *The 14th Human Genome meeting (HUGO)*, Montpellier, France, May 2010:
  Johannes F, **Wardenaar R**, Colomé-Tatché M, Mousson F, de Graaf P, Mokry M, Guryev V, Timmers HT, Cuppen E, Jansen RC. Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq.

- *Benelux Bioinformatics Conference*, Maastricht, The Netherlands, December 2008:
  Johannes F, **Wardenaar R**, Jansen RC. Mapping chromatin polymorphisms using ChIP-chip.

Awards

- Tweede landelijk Wiskunde B dag (Second national mathematics day) 2001, Second best mathematics report (High school)

- KNCV Gouden spatel (Royal Dutch Chemical Society; golden spatula) 2009, Second best undergraduate thesis